Statistical Methods for Social Scientists

Statistical Methods for Social Scientists

An Introduction

Lillian Cohen

Bureau of Social Science Research
The American University

Prentice-Hall of India Private Ltd.

Rs. 7.50

PRENTICE-HALL INTERNATIONAL, INC. Englewood Cliffs
PRENTICE-HALL OF INDIA, PVT. LTD. New Delhi
PRENTICE-HALL INTERNATIONAL, INC. London
PRENTICE-HALL OF AUSTRALIA, PTY. LTD. Sydney
PRENTICE-HALL OF CANADA, LTD. Toronto
PRENTICE-HALL FRANCE, S.A.R.L. Paris
PRENTICE-HALL OF JAPAN, INC. Tokyo
PRENTICE-HALL DE MEXICO, S.A. Mexico City

Reprinted in India by special arrangement with
PRENTICE-HALL, INC. Englewood Cliffs, N.J., U.S.A.

By PRENTICE-HALL OF INDIA (PVT.) LTD.

This book has been published with the assistance of the Joint Indian-American Standard Works Programme.

Printed by C. L. Bhargava at G. W. Lawrie & Co., Lucknow, and published by Prentice-Hall of India (Pvt.) Ltd., New Delhi.

1016-E

PREFACE

This book is designed as an introduction to statistics for social scientists. It is comprehensive enough to give insight into the logic involved in statistical manipulation, yet simple enough to be understood by anyone who has taken an elementary algebra course. If you can pass the introductory test in mathematics at the end of Chapter 1, you should have no difficulty understanding the techniques used in the book.

In order to facilitate the use of statistical technique in the analysis of social science data, most of the topics begin with a social science problem, and new concepts and techniques are introduced in solving the problem. Unless the source is specified, the data are hypothetical.

Problems from recent quantitative research findings in sociology and peripheral fields are included in the exercises. Examples are selected where statistical techniques are used in a theoretical framework.

I am indebted to Professor Sir Ronald A. Fisher, F.R.S. Cambridge, to Dr. Frank Yates, Sc.D., F.R.S. Rothamsted, and to Messrs. Oliver and Boyd Limited, Edinburgh, for permission to reprint Tables III, IV, and XXXIII, (I) and (II), from their book *Statistical Tables for Biological, Agricultural, and Medical Research.*

I also feel a very personal indebtedness to Robert Weiss of the Survey Research Center, University of Michigan, who read the manuscript several times; his insightful comments were of tremendous help. Any errors are mine, but many of the constructive revisions are his.

LILLIAN COHEN

CONTENTS

1. INTRODUCTION 1

1.1. The Problem, *2*

1.2. The Methodological Techniques in Answering the "Why" Question. The Hypotheses, *2*

1.3. Operational Concepts, *4*

1.4. Sampling Procedure, *5*

1.5. Statistical Analysis, *5*

1.6. Deviant Case Analysis, *6*

1.7. Conclusions, *6*

2. CONDENSING THE DATA IN TABULAR AND GRAPHIC FORM 9

2.1. Condensing the Data in Tabular Form, *9*

2.2. Condensing the Data in Graphic Form, *21*

3. MEASURES OF CENTRAL VALUE AND DISPERSION . 34

3.1. Measures of Central Value, *34*

3.2. Measures of Dispersion, *44*

4. USE OF THEORETICAL DISTRIBUTIONS AS MATHEMATICAL MODELS 51

4.1. The Normal Curve, *51*

4.2. The Binomial Distribution, *63*

5. SAMPLE DESIGN 74

5.1. Probability Sampling, *74*

5. SAMPLE DESIGN (*cont'd*)

5.2. What Kinds of Errors Can Be Made in Generalizing from a Sample to a Universe, *79*
5.3. Quota Sampling, *82*

6. INTRODUCTION TO STATISTICAL INFERENCE 85

6.1. Sampling Distributions, *85*
6.2. Two Methods of Statistical Inference, *89*

7. STATISTICAL INFERENCE CONTINUED 94

7.1. The Mean: Testing Hypotheses and Making Estimations, *94*
7.2. Testing Hypotheses about the Difference between Two Means, *102*
7.3. The *t*-Test of Significance for Small Samples, *105*
7.4. A Proportion: Testing Hypotheses and Making Estimations, *109*
7.5. Testing Hypotheses for the Difference between Two Sample Proportions, *116*
7.6. The Chi-Square Test of Significance, *120*

8. MEASURES OF ASSOCIATION 129

8.1. Association of Discrete Variables, *129*
Four-fold table analysis, *129* · A measure of the degree of association, *134*
8.2. Association of Continuous Variables, *139*
Linear regression, *139* · Linear correlation, *147* · Computation of correlation coefficient for grouped data, *152* · Questions to ask about a coefficient of correlation, *153*

APPENDIX 161

INDEX 177

Statistical Methods for Social Scientists

CHAPTER 1

INTRODUCTION

We shall regard *statistics* as the branch of scientific method that deals with the collection, description, and analysis of data whose occurrences or measurements have been counted.

The study of statistics has many different facets. We may want to order and describe data collected from a nearby industrial plant. Information might be collected on the marital status of all the industrial workers in the plant and on the relationship between their marital status and job satisfaction. We are interested solely in describing the workers of this plant, and not in generalizing our findings to a larger universe.

We may want to regard our data as a sample of a specified universe and infer characteristics of the universe, on the basis of the sample, with a certain degree of probability. We know the number of people who attend church five different Sunday mornings, for example, and want to estimate the average Sunday morning church attendance throughout the year. We know how a five per cent sample of students from a certain college is going to vote in the next election and want to predict how the total college population will vote.

Again, we may want to test the hypothesis that there is no difference in the proportion of churchgoers between Catholics and Protestants against the alternative hypothesis that Catholics are the more frequent churchgoers.

For most social science problems, we cannot set up a perfect laboratory situation, as is generally done in the physical sciences where all relevant factors may be controlled except the one factor allowed to vary. Most of the data that the social scientist deals with are observational rather than experimental. Observations are made in the real world and not in an experimental one. The science of statistics provides techniques for artificially controlling factors when we are dealing with observational data.

Statistics provides new ways of looking at data and new means of manipulating data. But it must be applied to have meaning. It is method, not content, but in any scientific study, method cannot be divorced from content. Hence, in learning research methods used by the social scientist, we should simultaneously learn about the content of research studies in the social sciences. As an illustration of the use of statistics in a social science context, we shall cite the Bettelheim and Janowitz study, *Dynamics of Prejudice*.[1]

1.1. The Problem

Scientific method is used within the context of a problem-solving situation. We have a problem requiring a solution or a question requiring an answer. We might ask the sort of question that can be answered by a simple descriptive statement: what proportion of veterans in Chicago took a post-war job at a lower occupational classification than their pre-war job? Our methodological techniques in solving this problem involve four steps. (1) We define the term "lower classification." (2) We delimit the universe. Does the universe of veterans, for example, include only those who got their discharge papers in Chicago, or does it include all those veterans living in Chicago at the time of the study? (3) We collect the required facts. The data may be available in the area office of the United States Bureau of the Census, in the local Veterans Administration Office, or in some local fact-finding agency. (4) We order and present the data in an informative table.

Instead of asking a question that can be answered with a simple description, we might have asked a "why" question: why is there intolerance against the Jews among some Chicago veterans and not among others? What are the factors associated with anti-semitism?

1.2. The Methodological Techniques in Answering the "Why" Question. The Hypotheses

We set up an exploratory *hypothesis* as an initial step in answering the question. It is an informed hunch, a tentative answer. It helps to focus our attention in selecting observations. We do not choose observations at random from the infinite number of possible obser-

[1] Bruno Bettelheim and Morris Janowitz, *Dynamics of Prejudice* (New York: Harper & Bros., 1950).

vations. Our hypothesis: those veterans who experienced downward mobility are more likely to be intolerant; those who were upwardly mobile are more likely to be tolerant. Intolerance against the Jew is related to the individual's *mobility* within the social structure. It is related more to his social mobility than to his economic or social position or his political or religious attitude at any one time.

The hypothesis is a statement of relationship between the characteristics "intolerance against the Jew" and "social mobility."

Hypotheses may be gotten from previous knowledge of subject matter or theory. They may be deduced from propositions which have or have not been validated; the testing of the hypothesis helps to validate the basic theory. The hypotheses in the Bettelheim and Janowitz study are based on psychoanalytic theory.[2]

We want to translate the exploratory hypothesis derived from the theory into two statistical hypotheses: the null hypothesis and the alternative hypothesis. Our *null* hypothesis states that the random samples of upwardly and downwardly mobile veterans come from universes where there is *no difference* in the intolerant proportion. We call this a *null* hypothesis because it is an hypothesis of no difference or no effect. We want to test this hypothesis of no difference against the *alternative* hypothesis that there is a difference greater than zero between the intolerant proportion of upwardly mobile and downwardly mobile veterans.

We shall adopt a procedure to choose between the null hypothesis and the alternative hypothesis. If the difference in intolerant proportions between the two samples lies in what we have previously decided is our region of acceptance, the null hypothesis will be validated. If the difference in intolerant proportions between the two samples lies in what we have previously decided to be the region of rejection, we shall reject the null hypothesis in favor of our alternative hypothesis. The statistical test is a formal technique for choosing between hypotheses. However, we cannot actually prove by statis-

[2] Other basic hypotheses in the Bettelheim and Janowitz study: ethnic intolerance is a function of the hostile individual's feeling that he has suffered deprivations in the past and a function of his anxiety about the future. Feeling deprived in the past and anxious about the future, and lacking ego strength and adequate controls, the intolerant person projects undesirable characteristics that he denies in himself to members of the outgroup.

We are primarily concerned here with the social and economic rather than the psychological correlates of intolerance studied in *Dynamics of Prejudice*.

tical test that an hypothesis is true, since we have only a sample and not a universe of observations.[3]

1.3. Operational Concepts

The statistical hypotheses have to be expressed in more specific form before testing. We want observable indices for the characteristics, intolerance against the Jew and social mobility.

Anti-Semitic Intolerance. After four to seven hours of intensive interviews, given by trained psychiatric social workers, each veteran was classified into one of four classes on an anti-Semitic continuum.

The Anti-Semitic Continuum

	(I) *Intensively Anti-Semitic*	(II) *Outspokenly Anti-Semitic*	(III) *Stereotyped Anti-Semitic*	(IV) *Tolerant*
With regard to restrictive action:	Spontaneously outspoken for restrictive action against Jews	Outspoken hostility shown after direct questioning	No expression of desire for hostile or restrictive action against Jews	
With regard to stereotyped opinions:	Wide range of unfavorable stereotyped opinions about Jews		A variety of stereotyped notions about Jews	No elaborate stereotyped beliefs about Jews. Isolated beliefs

The stereotypes about Jews represent them, for the most part, as a powerful well-organized group which, by inference, threatens the individual.

Social Mobility. A shift upward of one or more grades on the Edwards' socio-economic scale was considered upward social mobility; a reverse shift was classified as downward mobility.[4]

The definitions of "intolerance" and "social mobility" should meet certain criteria of fitness: the criteria of reliability and validity.

[3] The concepts *null hypothesis* and *alternative hypothesis* and *regions of acceptance and rejection* will be reintroduced in Chapters 6 and 7.

[4] Alba Edwards, *A Social-Economic Grouping of Gainful Workers of the United States* (Washington: Government Printing Office, 1938).

The definition is reliable if it enables experienced interviewers, using the same definition, to categorize people in the same way, and if it enables a single experienced interviewer, categorizing the same people at different times, to give the same classification for each person.

The definition is valid if we have really measured what we want to measure, that is, if we really mean by social mobility, for example, the movement up or down on the Edwards' socio-economic scale.

1.4. Sampling Procedure

To get a random sample of Chicago veterans, every ninetieth case was examined among the 15,000 discharge papers registered with the Recorder of Cook County, which includes the city of Chicago. If the case was a male enlisted veteran of World War II, under 35 years of age, who had been discharged between August and November 1945, and if he was neither Negro, Jewish, Chinese, Japanese, nor Mexican, he was included in the sample. An attempt was made to interview everybody selected for the sample. Total compliance failed by about 14 per cent mostly because of movement to other areas and refusal to be interviewed. One hundred and fifty veterans were interviewed.

1.5. Statistical Analysis

Social and economic characteristics of the veterans were studied in an exploratory manner to determine whether there was an association with anti-Semitism. The characteristics included age, educational level, religious denomination, political affiliation, family composition, nativity of parents, socio-economic status, and reading and listening habits. There appeared to be no very significant association between these characteristics and anti-Semitism.

The concepts of status and status characteristics were replaced by the concept of social mobility. Men were asked about their occupational status before the war and at the time of the interview (about six months after their discharge). The findings are given in Table 1-1. A statistical analysis showed that the *difference in intolerant proportions* between the upwardly and downwardly mobile sample of veterans lay in the "region of rejection" of the null hypothesis, and was more compatible with the alternative hypothesis. Consequently, the null hypothesis of no difference in intolerant proportion between

upwardly and downwardly mobile veterans *was rejected* in favor of the alternative hypothesis that veterans who experienced downward mobility were likely to have a higher intolerance rate than upwardly mobile veterans.

Table 1-1. Anti-Semitism and Social Mobility among 130 World War II Veterans, Chicago, 1946

	Downward Mobility		No Mobility		Upward Mobility		Total	
	No.	%	*No.*	%	*No.*	%	*No.*	%
Tolerant	2	11	25	37	22	50	49	38
Stereotyped	3	17	26	38	8	18	37	28
Outspoken and intense	13	72	17	25	14	32	44	34
Total	18	100	68	100	44	100	130	100

Source: Bettelheim and Janowitz, *Dynamics of Prejudice* (New York: Harper and Bros., 1950), page 59.

1.6. Deviant Case Analysis

Deviant cases consisted of those veterans who did not follow the general trend, but were, for example, upwardly mobile and outspokenly intolerant. They were examined to find reasons for deviation. It was found that the outspokenly intolerant group of upwardly mobile veterans was considerably more mobile than the others. An explanation tentatively given for their deviation was that sharp upward mobility was likely to be associated in general with marked aggressiveness. It implied changes in life patterns which produced great stress in the individuals.

1.7. Conclusions

Bettelheim and Janowitz concluded that to understand intolerance among the sampled population it was less important to concentrate on the social and economic background of the individual than to investigate the nature of his social mobility. The results of the study are generalizable to other populations only if the Chicago World War II veterans are representative of other populations in characteristics relevant to intolerance and social mobility.

The procedures of all social science studies do not follow chronologically the order outlined in the *Dynamics of Prejudice* study. Reichenbach has made the useful distinction between the "context of

discovery" and the "context of justification."[5] The context of discovery refers to the psychological processes involved in achieving insight. Insight is a sine qua non of any worthwhile study, and at every stage of the study. But we have little to say here about what makes a person an insightful scientist. The context of justification yields the proof of the hypothesis, that is, the series of logical operations justifying the conclusions. It is within the context of justification that the studies are presented in succeeding chapters.

KEY TERMS

deviant cases **statistical hypothesis**
reliability **validity**

REFERENCES

Bettelheim, Bruno, and Morris Janowitz, *Dynamics of Prejudice.* New York: Harper and Bros., 1950.

Cohen, Morris R., and Ernest Nagel, *An Introduction to Logic and Scientific Method.* New York: Harcourt, Brace and Co., 1934.

Goldhamer, Herbert, "An Outline of Social Science Methodology." Unpublished manuscript, The University of Chicago, 1947.

Merton, Robert, "Sociological Theory," *American Journal of Sociology,* Vol. 50, (May 1945), 462–73.

INTRODUCTORY EXERCISE IN MATHEMATICS TO TEST THE USE OF SIMPLE ARITHMETIC AND ALGEBRAIC SYMBOLS

Mark each problem either true or false. If false, give the correct answer.

(1) $(\frac{2}{4})^2 = \frac{25}{4}$

(2) $\frac{10}{5} = 3\frac{1}{5}$

(3) $\frac{12}{2+5} = 8\frac{2}{5}$

(4) $\frac{12}{2 \times 5} = 1\frac{1}{5}$

(5) $(\frac{2}{3})(2+4) = \frac{8}{3}$

(6) $(.3)(\frac{1}{3}) = .01$

(7) $\frac{1}{5}(1 - \frac{2}{5}) = -(\frac{1}{5})^2$

(8) $(\sqrt{a})^2 = a$

(9) $\sqrt{9+81} = \sqrt{9} + \sqrt{81}$

(10) $\frac{6+5}{6} = 1 + \frac{5}{6}$

(11) $.001 \div .01 = .1$

(12) $.5 \times .05 = .25$

(13) $\frac{99}{1000} = 99\%$

(14) $88\frac{1}{3}\% = \frac{5}{6}$

(15) $\sqrt{.25} = .5$

(16) $\sqrt{.025} = .05$

[5] Hans Reichenbach, *Elements of Symbolic Logic* (New York: The Macmillan Company, 1947), p. 2.

(17) $\sqrt{6400} = 10\sqrt{64}$

(18) $1.01 + .001 = 1.02$

(19) $(.06)(4.25) = .00255$

(20) $\frac{1}{3} + \frac{1}{3} - \frac{1}{6} = \frac{2}{3}$

(21) $.004 = .4\%$

(22) $\sqrt{9 \times 81} = \sqrt{9} \times \sqrt{81}$

(23) $\frac{5}{10} = .50$

(24) $(\frac{1}{2})(25\%) = .125\%$

(25) $\sqrt{4000} = 20$

(26) $(\frac{1}{2})(\frac{2}{4}) = (\frac{1}{4})(\frac{2}{2})$

(27) $\frac{6 \times 5}{6} = 1 \times \frac{5}{6}$

(28) $3! = 3 \times 2 \times 1$

(29) $3^3 = 30$

(30) $\sqrt{9^2} = 9$

(31) $.70 = \frac{7}{10}$

(32) $(a + b)^2 = a^2 + b^2$

(33) In a/b, a is the numerator and b, the denominator.

(34) In the terms $a^2 + 2ab + b^2$, the coefficient of the first and last term is 1; the coefficient of the middle term is 2.

(35) In the terms $a^2 + 2ab + b^2$, there are three exponents of the second degree.

(36) In the equation $Y = 20 + 3X$, when $X = 2$, $Y = 26$.

(37) If $r^2 = 1 - \frac{9}{25}$, $r^2 = .64$ and $r = .08$.

(38) If $r^2 = .81$, $r = .90$

(39) $\frac{6!}{4!} = 30$

(40) In the equation $Y = 4 - 2X$, when $X = \frac{1}{2}$, $Y = 2$.

CHAPTER 2

CONDENSING THE DATA IN TABULAR AND GRAPHIC FORM

In this chapter we attempt to answer three questions concerning the more than 150 million people of the United States: (1) whether they live in urban or rural areas, (2) whether they are male or female, and (3) how old they are. After we have collected the information, which has already been classified by the United States Census, we shall systematize our results by presenting the data in tabular and graphic form.

The process of classifying people by some observable characteristics is done all the time. We say that among the ten families who live in our city block, three live in modern ramblers, five in Cape Cod colonials, and two in homes of Spanish-style architecture. Or, we might say that among the ten students in a statistics class, five live within a radius of one mile from the university, three live one-to-two miles, and two, two-to-three miles. We have classified our data and enumerated the frequency for each class.

With a class of only ten students, we might have wanted to retain the original unclassified data of distance from school instead of classifying the data into one-mile intervals. The first student lives 1.2 miles from school, the second student, .5 miles, the third, 1.9 miles, etc. However, if our population consists not of ten students, but of a much larger universe, some classification is crucial if we are to comprehend the significance of the information.

2.1. Condensing the Data in Tabular Form

Problem 1

It has been estimated that nearly three-fourths of the world population is rural. We want to find out how the United States compares

with the rest of the world in this regard, i.e., how much of the population of the United States in 1950 lived in rural-nonfarm and in rural-farm areas, and how much lived in urban areas.

There are limitations inherent in a statistical comparison. The rural-nonfarm population within the social and economic confines of a large urban center is not rural in the same sense as the rural-nonfarm population of the wide open spaces.[1] Similarly, farmers in rural-farm sections of the United States who have radios, television sets, and daily newspapers are pervaded by urban influences much more than are isolated peasants in some countries of Europe and Asia.

Tabular Presentation of Data. A One-Way Table. In the first column of a table we state the characteristic being tabulated, and underneath, we list its categories. The categories must satisfy three criteria. (1) They must be exhaustive, enabling each of the 150 million inhabitants of the United States to fit into *one* of the categories. (2) They must be mutually exclusive, i.e., one category does not overlap another. Each of the inhabitants can fit into *only one* category. (3) They must be relevant to the problem—in our example, the problem of determining what proportion of the United States population is rural.

Urban-rural residence is the characteristic of the tabulation in Table 2-1; its two categories are urban population and rural population. Within rural population there are two subcategories: rural-nonfarm and rural-farm.

In the second column of Table 2-1 is given the number of cases in each category (the frequency for that category), and in the third and

[1] In its 1950 tabulations, the United States Bureau of the Census revised its definitions of "urban" and "rural" to help take account of this limitation. The following are its new definitions.

(1) Urban population consists of all persons living in: (a) Places of 2,500 inhabitants or more incorporated as cities, boroughs, villages, and towns (some exceptions are noted to the town classification); (b) The densely settled urban fringes incorporated or unincorporated, around cities of 50,000 or more; and (c) Unincorporated places of 2,500 inhabitants or more outside any urban fringe.

The urban population in the United States for 1950 is about 8 million greater than it would have been under the definition used in the 1940 census.

(2) Rural-nonfarm population consists of persons living in rural territory, but not on farms.

(3) Rural-farm population consists of all rural residents living on farms.

SOURCE: Bureau of the Census, U.S. Department of Commerce, *1950 Census of Population, Advance Reports*, Series PC-14, No. 6, December 1, 1952.

fourth columns, the relative frequency in proportions and percentages.

From Table 2-1 we learn that about 54 million out of 151 million people in the United States in 1950 were rural inhabitants. The proportion of rural population to total population is 54,229,675 divided by 150,697,361, or .36. The proportion of urban population

Table 2-1. Population for the United States, Urban and Rural, 1950

		Relative Frequency*	
Urban-Rural Residence	*Frequency*	*Proportions*	*%*
(1)	(2)	(3)	(4)
Total population	150,697,361	1.000	100.0
Urban population	96,467,686	.640	64.0
Rural population	54,229,675	.360	36.0
Rural-nonfarm	31,181,325	.207	20.7
Rural-farm	23,048,350	.153	15.3

Source: Bureau of the Census, U.S. Department of Commerce, *1950 Census of Population, Advance Reports*, Series PC-14, No. 6, December 1, 1952.

* Ordinarily, either proportions or percentages, but not both, is presented in a single table. Both measures of relative frequency are given here for illustrative purposes.

to total population is .64. In contrast with much of the rest of the world, the population of the United States is more urban than rural.

If the *proportion* of urban population to total population is .640, then the *percentage* of urban population is .640 times 100, or 64.0.[2]

A Two-Way Table. Table 2-2 gives the sex distribution by urban-

Table 2-2. Population by Sex, for the United States, Urban and Rural, 1950

	Frequency (000's)			*Sex Ratio* (males per
Urban-Rural Residence	*Both Sexes*	*Male*	*Female*	100 females)
(1)	(2)	(3)	(4)	(5)
Total population	150,697	74,833	75,864	98.6
Urban population	96,468	46,892	49,576	94.6
Rural population	54,229	27,941	26,288	106.3
Rural-nonfarm	31,181	15,863	15,318	103.6
Rural-farm	23,048	12,078	10,970	110.1

Source: Bureau of the Census, U.S. Department of Commerce, *1950 Census of Population, U.S. Summary Bulletin*, Series P-B1.

[2] To change a proportion to a percentage, move the decimal point two places to the right.

rural residence. It shows how males and females are distributed among the urban and rural population. Whereas Table 2-1 is tabulated for only one characteristic, i.e., urban-rural residence, Table 2-2 is tabulated simultaneously for two characteristics: urban-rural residence and sex. By giving the joint distribution of *two* characteristics, Table 2-2 enables us to observe the *association* of these two characteristics as well as their individual distributions.

The sex ratio of 98.6 indicates that females exceeded males for the country as a whole in 1950.[3] Women were more plentiful than men among the urban population, but in the rural areas males outnumbered females. The sex ratio implications are important in such diverse areas as marriage patterns, women's vote, consumer spending, and labor force characteristics.

Table Construction. Examine the parts of Tables 2-1 and 2-2. The five essential parts are:

(1) Table heading: (a) table number; (b) title—tells what? where? when?

(2) Heading for horizontal rows of data (the left-hand column and its heading).

(3) Heading for vertical columns of data.

(4) Body of statistical data.

(5) Footnotes and sources.

Every table should be sufficiently complete to stand alone without reference to the text, although it may be accompanied by an interpretation in the text. The total may appear at the top or the bottom of a table. It is set off from the rest of the table by a space or line.

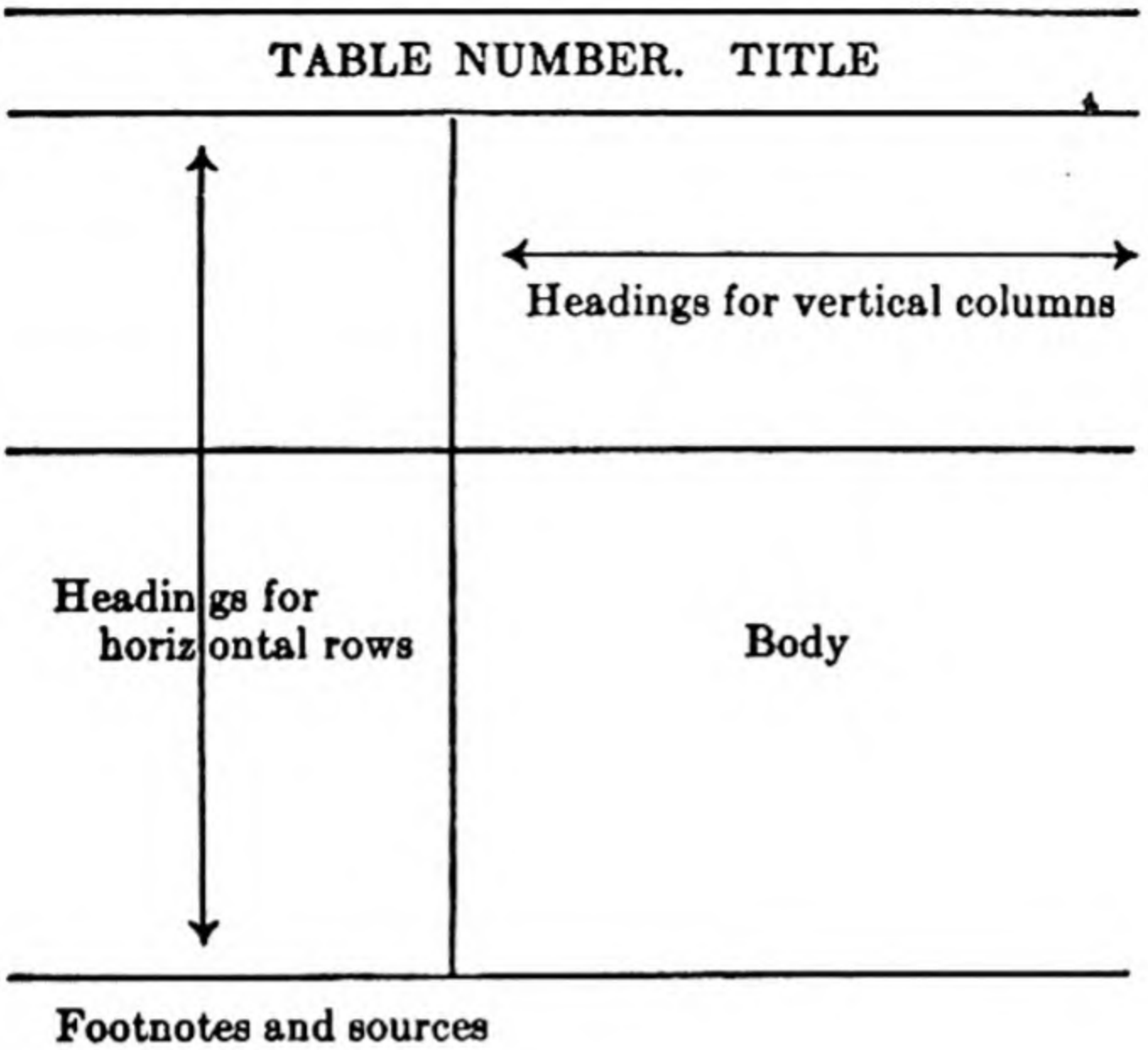
In any table where percentages or proportions are given without the corresponding numbers, the base number (the denominator) should be indicated. In Table 2-1, the base is total population. Even if we were to omit the actual frequency figures (column 2) from Table 2-1, we would still retain the total population figure of 150,697,361, upon which the proportions and percentages are based.

If the population figures are rounded to the nearest thousand (total United States population, for example, being given as 150,697) we insert 000's or thousands under the appropriate heading.

[3] The sex ratio (column (5) of Table 2-2) is the number of males per 100 females.

Sex ratio, United States = (Total male population divided by total female population) times 100

$$(74,833,239 \div 75,864,122) \times 100 = 98.6$$



Problem 2

Changes in age distribution within the United States affect many social areas such as the number of schools, the training of teachers in the necessary grades, and the provision of old-age pensions. We want to examine recent changes in the United States age distribution.

Tabular Presentation of Data. The age classification is based on the age of the person at his last birthday before the date of the census, that is, his age in completed years.

If we are interested in the age distribution of the population in 1940 and in 1950, we use total population in each of these years as the base from which to compute proportions or percentages. (Table 2-3, cols. 4 and 5.)

Age	*1940*	*1950*
Total	100.0	100.0
Under 5 years	8.0	10.7
5 to 9	8.1	8.8

If we are primarily interested in the trend over time, we use as our base the 1940 population for each age group and determine the change over the decade. To compute proportionate change: (1) Determine the numerical increase or decrease over the period. (2) Divide the amount of change by the original amount (the base). The numerical

increase in children under 5 years of age from 1940 to 1950 (5,622,000) is divided by 1940 population (10,542,000) to give the proportionate change over the period. The *proportionate* change is .533, the *percentage* change is 53.3. (Table 2-3, col. 6.)

Table 2-3 gives the age distribution, in percentages, for 1940 and 1950 (with total population for each year used as a base) as well as the age trend over the ten-year period (1940 population used as a base).

Table 2-3. Age Distribution for the United States, 1940 and 1950

Age	*Frequency (000's)*		*Relative Frequency (per cent)*		*Per Cent Change, 1940 to 1950*
	1940	*1950*	*1940*	*1950*	
Total	131,669*	150,697	100.0	100.0	14.5
Under 5 years	10,542	16,164	8.0	10.7	53.3
5 to 9	10,685	13,200	8.1	8.8	23.5
10 to 14	11,746	11,119	8.9	7.4	−5.3
15 to 19	12,334	10,617	9.4	7.0	−13.9
20 to 24	11,588	11,482	8.8	7.6	−0.9
25 to 29	11,097	12,242	8.4	8.1	10.3
30 to 34	10,242	11,517	7.8	7.6	12.4
35 to 39	9,545	11,246	7.2	7.5	17.8
40 to 44	8,788	10,204	6.7	6.8	16.1
45 to 49	8,255	9,070	6.3	6.0	9.9
50 to 54	7,257	8,272	5.5	5.5	14.0
55 to 59	5,844	7,235	4.4	4.8	23.8
60 to 64	4,728	6,059	3.6	4.0	28.2
65 to 69	3,807	5,003	2.9	3.3	31.4
70 to 74	2,570	3,412	2.0	2.3	32.8
75 and over	2,643	3,855	2.0	2.6	45.9

SOURCE: Bureau of the Census, U.S. Department of Commerce, *1950 Census of Population, Advance Reports*, Series PC-14, No. 5, October 31, 1952.

* The difference between the sum of the frequencies and the total is due to rounding.

According to Table 2-3, the percentage of the population in the youngest and oldest groups is increasing relative to the other groups. The census speculates on some of the reasons for this increase at the youngest and oldest ages as follows: an increase in the birth rate in the 1940's, a decline in infant mortality, and a longer life expectation.

Discrete and Continuous Variables. The characteristics studied in Tables 2-1 and 2-2, urban-rural residence and sex, are discrete variables.

A *variable* is a characteristic that can take on more than one value.

A *discrete variable* has values that vary either (1) qualitatively, or (2) by integral amounts, e.g., 1, 2, 3, etc.

Sex is a qualitative discrete variable. It has the values male and female. Number of children per family is a nonqualitative discrete variable. It can take on the values 0, 1, 2 . . . , but not the values .1, .2, . . . 1.1, 1.2, ...

Table 2-3 refers to age, which is a *continuous variable*, able to take on any value (integral and fractional) within a certain range. A person can be five years old, 5 years and 1 day, 5 years and 2 days, etc. When age is defined in terms of a person's last birthday, the continuous variable age is recorded as a discrete variable. A person is 5 years old, or 6 years, or 7.

Choice of Intervals. The variable age is classified by the United States Census into five-year age intervals throughout most of Table 2-3. If the age variable had been divided into one-year groups, there would have been more than 75 divisions, resulting in an unwieldy classification. If the variable had been divided into twenty-year age intervals, the shape of the distribution might have been concealed. The number of intervals is an arbitrary choice, depending upon the desired use of the data—whether, for example, great precision or an over-all picture is the prime requisite. It is customary to make the size of an interval sufficiently large to permit from 10 to 25 intervals. With less than ten intervals, too much precision is often lost; with more than 25 the classification becomes cumbersome. The age variable in Table 2-3 has 16 intervals.

In grouping people into 5-year age intervals, the identity of the original ages is lost. We do not know how the ages are distributed within the interval itself. But we do get a picture of the age distribution over the entire range, from lowest to highest age.

Intervals of equal size are desirable, since they facilitate computations and comparisons. However, it may be necessary to condense the table to save space and money, prevent unwieldiness, or conceal the identity of a few cases. In these instances, there may be open-end intervals such as 75 years and over in Table 2-3, where it is impractical to continue the 5-year groups for the increasingly small percentage of cases. On the other hand, it may be necessary to expand part of the table because of the need for more precise information in certain groups. In an age distribution table with 5-year intervals, 1-year intervals may be wanted, for example, under the age of 5.

When we are dealing with a continuous variable, we must decide what kind of interval we want to work with. The *real* limits are not automatically given when we specify two integral values. For example, the second interval in Table 2-3, described by the integers 5 to 9, will generally have one of two possible sets of *real* limits, with consequent differences in midpoints.

REAL INTERVALS:

	Lower Limit	*Midpoint*	*Upper Limit*	
(1)	4.50	7.00	9.49	(i.e., up to, but not including 9.50)
(2)	5.00	7.50	9.99	(i.e., up to, but not including 10.00)

Both of these real intervals are *described* by the closest integral values, 5 to 9.*

A procedure for selecting real intervals involves the following:

(1) The determination of the range of the variable, that is, the difference between the highest and lowest value. This working range is modified where open-end intervals are necessary.

(2) The division of the range into 10 to 25 equal intervals of some convenient size, in many cases 5 or 10 units.

(3) The selection of the interval midpoint, wherever possible, at the point where cases tend to cluster, since the midpoint is often used to represent all the cases in the interval.

(4) The selection of interval limits, wherever possible, at points that make it easy to sort cases into intervals. If the limits come at points where cases cluster, we must make an arbitrary decision about what to do with the cases falling on the limits.

There is obviously no unique choice of interval limits or midpoints for given data.

Choice of the Base for Percentages. We shall often want to relate one variable to another in percentage terms. To do this, we must decide upon the base for the percentages. In order to determine the

* In classifying the population by age, the census did not have the problem of determining which real interval to apply to a continuous variable, since "age" was recorded in integral values by age as of last birthday. The real interval for 5 to 9 would be 5.00 to 9.99.

base, we must know what we want to find out. In Table 2-4, for example, are we interested in discovering (1) the proportion of intolerant veterans who are downwardly mobile; or (2) the proportion of downwardly mobile veterans who are intolerant?

Table 2-4. Anti-Semitism and Social Mobility among 130 World War II Veterans, Chicago, 1946

	Downward Mobility	*No Mobility*	*Upward Mobility*	*Total*
Tolerant	2	25	22	49
Stereotyped	3	26	8	37
Outspoken and intense	13	17	14	44
Total	18	68	44	130

SOURCE: Bettelheim and Janowitz, *Dynamics of Prejudice* (New York: Harper and Bros., 1950), page 69.

Our hypothesis of Chapter 1 states that intolerance is related to an individual's mobility within the social structure. The level of tolerance is considered to depend upon mobility. Hence mobility is regarded as the independent variable.

Unless other information is desired, it is customary to use as the base for percentages the variable which can be regarded as the independent variable; in the intolerance-mobility example, it would be mobility and not level of tolerance. Hence our percentages would be computed vertically, with downward mobility (18 cases), no mobility (68 cases), and upward mobility (44 cases) used as the bases.

KEY TERMS

base for percentages
continuous variable
discrete variable
frequency
percentages
proportions
relative frequency
real interval variable

REFERENCES

Hagood, M. J., and D. O. Price, *Statistics for Sociologists*, chap. 5. New York: Henry Holt and Company, 1952.

Lindquist, E. F., *A First Course in Statistics*, chap. 2. Boston: Houghton Mifflin Company, 1942.

Treloar, Alan E., *Biometric Analysis*, chap. 3. Minneapolis: Burgess Publishing Company, 1951.

United States Department of Commerce, *Bureau of the Census Manual of Tabular Presentation*. Washington, D.C.: Government Printing Office, 1949.

EXERCISES

1. (a) From the United States Census of 1950, find out the per cent of all United States families by number of children per family under ten years of age. Prepare a table for these data, giving actual numbers and percentage distribution.

(b) What is the census definition of family? Would this definition be the same as a sociological definition?

(c) How many variables are there in the table? Discrete or continuous?

2. (a) Assume that you want to discover the home ownership rate by age of family head. Construct a blank table with the following parts: table heading, heading for horizontal rows and vertical columns. (To get a home ownership rate for each age group, we must know (1) the number of home owners in the age group and (2) the total number of home owners plus renters in that age group. (1) is the numerator and (2) is the denominator of the home ownership rate.)

(b) Now insert the statistical body of data from census material and add the source to the table.

(c) What is the census definition of "home owner"? How is age recorded in the census?

(d) What conclusion can be drawn about the relationship between home ownership and age of family head? How does this conclusion compare with Burgess and Locke's "life cycle of a family" from kitchenette apartment to larger apartment, to single home, a return to city apartment, and finally "refuge in an apartment hotel in old age"? (E. W. Burgess and H. Locke, *The Family*, page 520.)

3. Can a continuous variable be converted into a discrete variable? Can a discrete variable be converted into a continuous variable? Give examples where possible.

4. According to the table, what are the least desirable colors from the point of view of rural Negro youth? Speculate on sociological and psychological explanations for this percentage distribution.

Judgments of Rural Negro Youth on "Worst Color to Be" by Sex*
(In percentages)

Worst Color	*Both Sexes*	*Boys*	*Girls*
Black	34.9	34.5	35.1
Dark-brown	1.9	2.5	1.6
Brown	1.4	1.9	1.1
Light-brown	1.4	0.9	1.6
Yellow	28.2	28.8	27.9
White	32.1	31.3	32.6

SOURCE: Charles Johnson, *Growing Up in the Black Belt*. (Washington, D C.: The American Council on Education, 1941), p. 263.

* Based on responses of 635 boys and 1,161 girls.

5. The table cross-classifies the occupational level of Negro grooms with White brides.

Negro Grooms in Negro-White Marriages by Own Occupation and Occupation of Bride, 1914–1938
(Where both partners are gainfully employed)

Occupation of Negro Groom	Occupation of White Bride			
	I	II	III	
	Professionals, Proprietors, Managers, Clerks	*Skilled and Semi-skilled Workers*	*Unskilled Workers*	*Total Gainfully Employed*
I. Professionals, proprietors, managers, clerks	16	3	13	32
II. Skilled and semi-skilled workers	11	11	46	68
III. Unskilled workers	11	7	44	62
Total gainfully employed	38	21	103	162

Source: Louis Wirth and Herbert Goldhamer, "The Hybrid and the Problem of Miscegenation" in *Characteristics of the American Negro*, ed. Otto Klineberg (New York: Harper & Bros., 1944), p. 293.

(a) Do Negro groom-White bride marriages fit into the general category of marriages where there is a fairly close correspondence between the occupational levels of marriage partners?

What proportion of the Negro grooms married White brides in the same occupational category? (In what three cells do we find bride and groom in the same occupational category?)

What proportion of the Negro grooms in white-collar occupations (I) married White unskilled brides (III)?

What proportion of the unskilled Negro males (III) married White women in white-collar occupations (I)?

(b) White brides who marry Negro grooms tend to be at what occupational level?

6. In a study of two adjoining Massachusetts Institute of Technology housing projects peopled with married veterans who had little or no previous contact, it was hypothesized that friendships are likely to depend upon the ecological factor of physical distance. Data to support this hypothesis are given in the table.

Column 1 gives the approximate physical distance that can separate any two persons living on the same floor of any of the seventeen buildings of the Westgate West project.

Column 2 presents the total number of choices given to persons living on the same floor at each distance away from the chooser in answer to the question: What three people in the two projects do you see most of socially?

Column 3 gives the total number of possible choices at different units of physical distance. There are, for example, more one-unit-distance choices

The Relationship between Sociometric Choice and Physical Distance on One Floor of a Westgate West Building
(One of the housing projects)

Units of Approximate Physical Distance (1)	*Total Number of Choices Given* (2)	*Total Number of Possible Choices* (3)	*Choices Given (2) Divided by Possible Choices (3)* (4)
1	112	8 × 34* = 272	.412
2	46	6 × 34* = 204	.225
3	22	4 × 34* = 136	.162
4	7	2 × 34* = 68	.103

SOURCE: L. Festinger, S. Schachter, and K. Bach, *Social Pressures in Informal Groups* (New York: Harper and Bros., 1950), p. 38.

* There are 17 buildings, each with two floors.

than four-unit-distance choices possible on any single floor. Apartment [5] is 4 units away from apartment [1], 3 units away from [2], etc.

One Floor [1] [2] [3] [4] [5]

(a) Does this table substantiate the hypothesis of high relationship between friendships and physical distance? Explain.

7. The table gives the relationship between cohesiveness and prestige for

The Relation between Cohesiveness and Prestige in Nine Courts of the Westgate Housing Project

Court (1)	*Number of Residents* (2)	*Outside Choices Received Divided by Outside Choices Given* (prestige index) (3)	*Choices in Court Divided by Total Choices* (cohesiveness index) (4)	*Choices in Court Minus One-half Mutual Choices Divided by Total Choices* (revised cohesiveness index) (5)
Miller	13	.80	.56	.485
Carson	13	.81	.48	.403
Richards	7	.88	.47	.433
Freeman	13	1.06	.48	.419
Williams	13	1.06	.53	.447
Main	7	1.17	.67	.527
Howe	13	1.23	.63	.500
Rotch	8	1.30	.55	.523
Tolman	13	2.08	.62	.529

SOURCE: Festinger, Schachter, and Bach, *Social Pressures in Informal Groups* (New York: Harper and Bros., 1950), p. 97.

the nine courts of the Westgate project (one of the M.I.T. married-veterans' housing projects studied by Festinger, Schachter, and Bach).

The status or prestige position is measured by the ratio of friendship choices received in any housing court from outsiders to choices given to outsiders (column 3).

The cohesiveness of any court is measured by the percentage of friends that are in-court choices (column 4).

(a) Which courts had the highest prestige? Which, the greatest cohesiveness? What appears to be the relationship between the cohesiveness of a court and its status in the social structure of the housing project?

(b) Within any court, there may be subgroups highly cohesive within themselves but without any friendship choices between the subgroups. We may take into account the subgroup formation by correcting for the number of mutual choices which occurred. Column (5) gives a corrected measure of cohesiveness; the proportion of in-court choices to total choices is corrected by subtracting from the in-court choices one-half the number of mutual choice pairs that occurred. Does the corrected measure of cohesiveness change the relationship between cohesiveness and social prestige?

2.2. Condensing the Data in Graphic Form

Graphic Presentation of the Variable Urban-rural Residence. One graphic illustration of the urban-rural distribution of United States population is a pie chart, the pieces of which add up to 100 per cent.

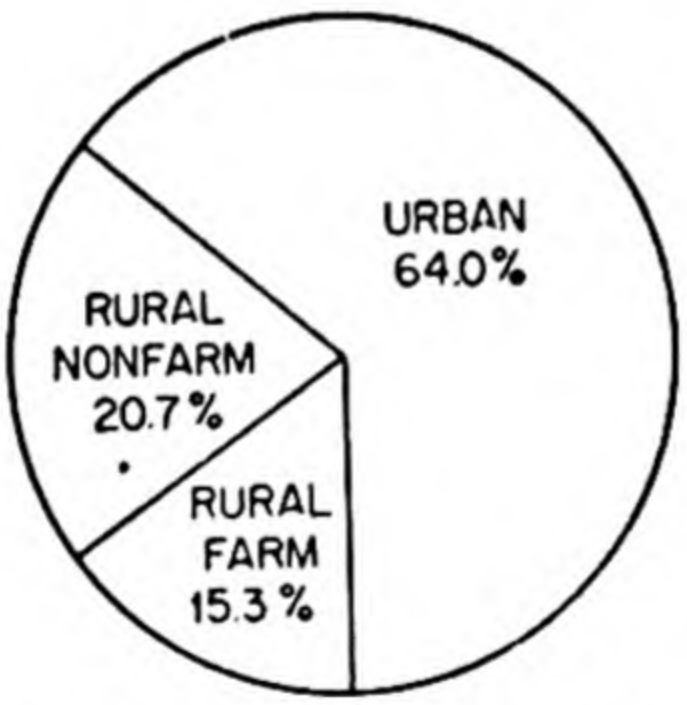


Fig. 2.1. Distribution of United States population by urban-rural residence, 1950. (Pie chart for data of Table 2-1.)

Another graphic illustration is the vertical bar graph. On the horizontal scale of the vertical bar graph are indicated the values of the variable urban-rural residence, on the vertical scale, the frequency or number of people, ranging from 0 to 100 million. The urban population is 96,467,686.

To draw the bars of the graph, a horizontal line is placed at a height corresponding to the population of the category. The horizontal line is closed by vertical lines erected at the category limits.

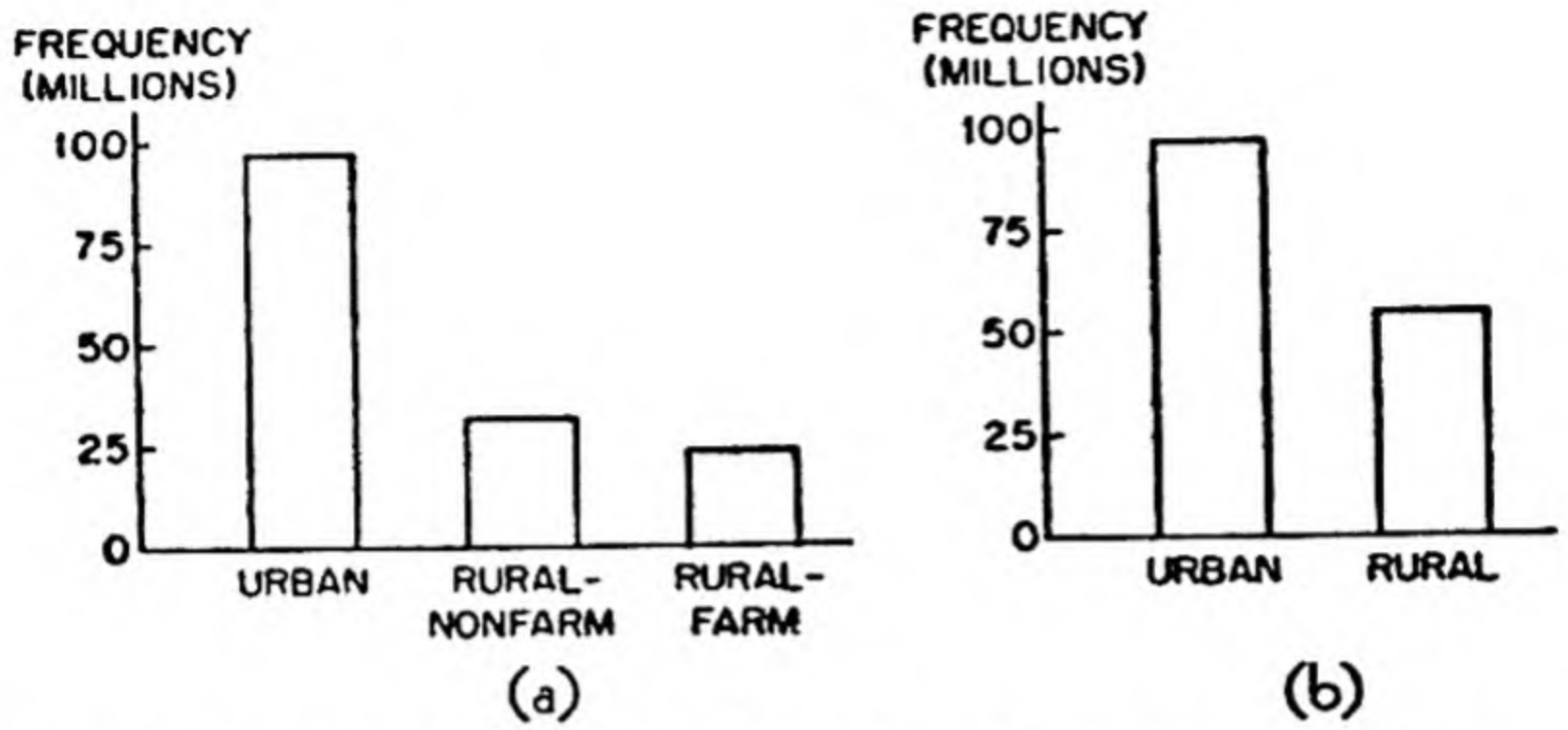


Fig. 2.2. United States population by urban-rural residence, 1950. (Graph for data of Table 2-1.)

The vertical scale always begins with zero. The heights of bars are proportional to the frequency, and proportions or percentages placed on the vertical scale give the same graph as do the absolute numbers. If the vertical scale does not begin with zero, the graph gives a distorted picture of the data. For example, in Fig. 2.3b, 15

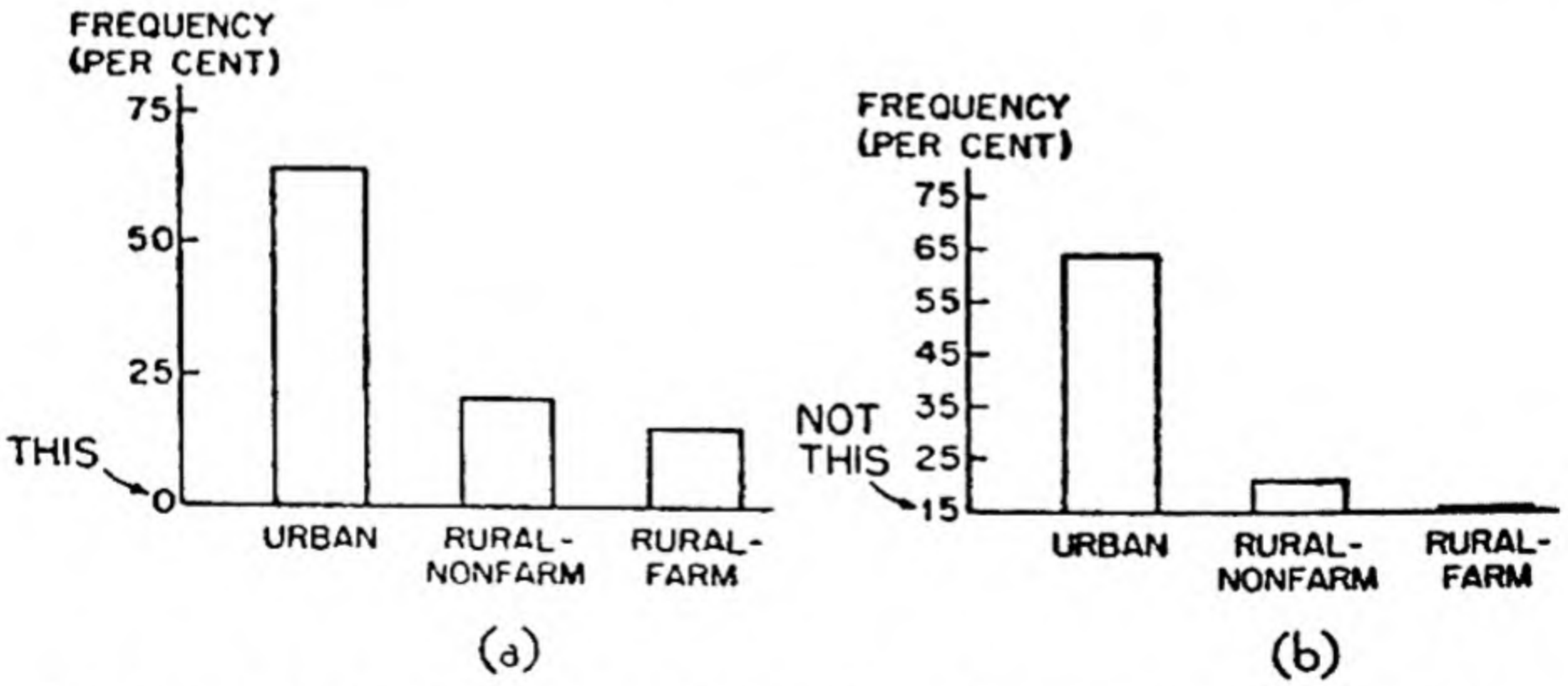


Fig. 2.3. United States population by urban-rural residence, 1950.

per cent is placed at the bottom of the vertical scale since no bar is lower than 15 per cent. But it appears from the size of the bars that

only a minute fraction of the United States population lives in rural-nonfarm and rural-farm areas, a fallacious interpretation.

Graphic Presentation of the Variable, Age. The frequency distribution for the variable, age, is shown graphically in Fig. 2.4 by a bar

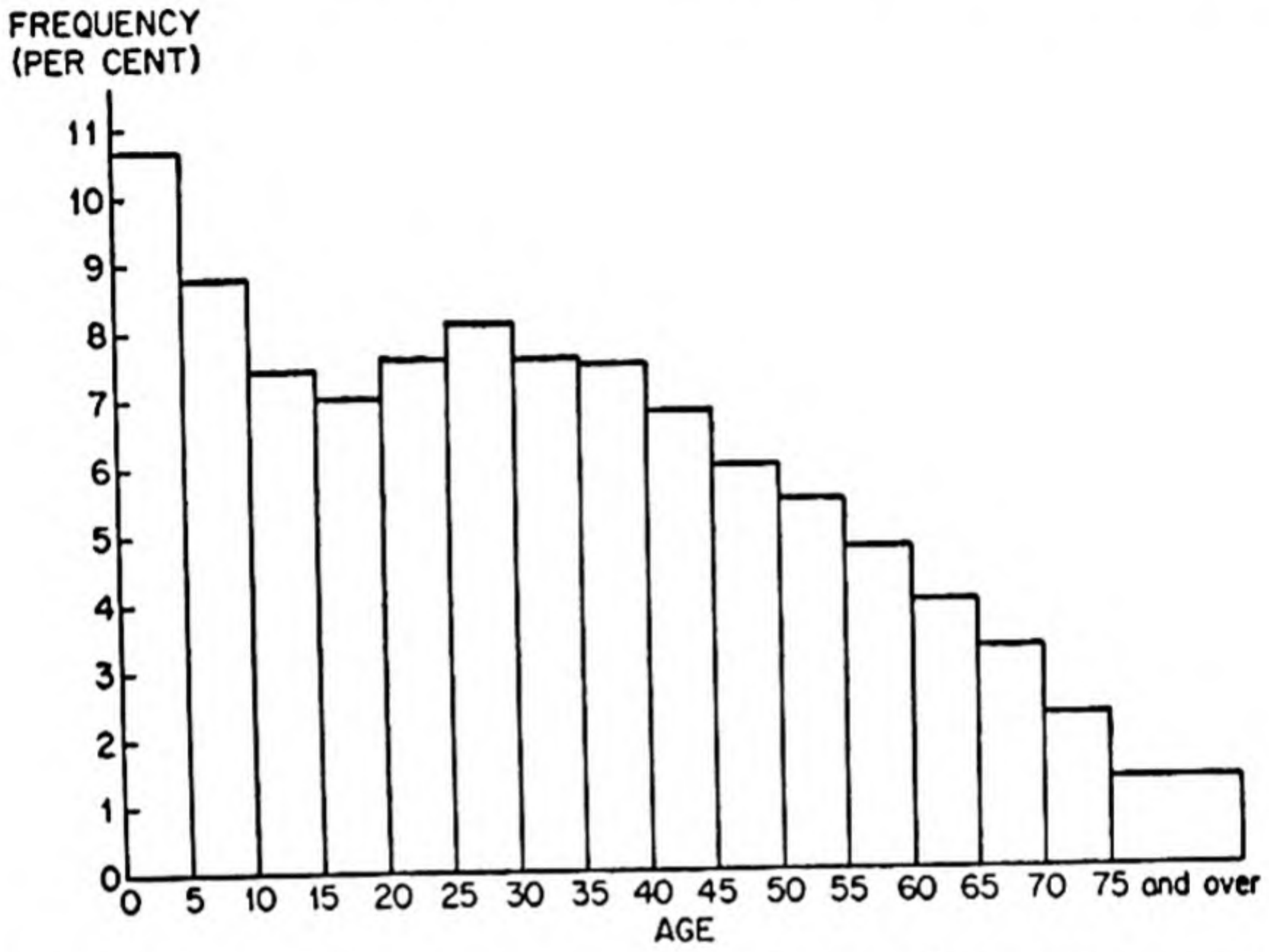


Fig. 2.4. Age distribution for the United States, 1950. (Histogram for data of Table 2-3.)

graph known as a histogram. The histogram is a graph whose bars give the frequency for a range of values, e.g., 5 to 10 years of age. Although age is a continuous variable, it is recorded by the Bureau of the Census in discrete units, i.e., age as of last birthday.

Along the horizontal scale of the histogram we mark off the limits of each age interval. Equal distances in age are given equal space along the scale. If the space between ages 5 and 10 is $\frac{1}{4}$ inch, then the space between ages 10 and 15 is $\frac{1}{4}$ inch.

Vertical lines are erected at the limits of the age interval. The first interval limits are 0 and 5, the second, 5 and 10. The vertical lines are closed to form rectangles at a height across from the frequency for that interval.

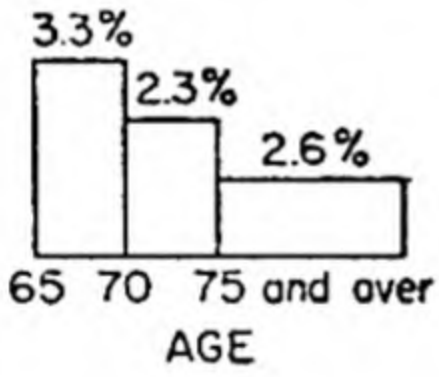
The *total area* inside all the rectangles of the histogram represents the total population. The area inside the first rectangle is propor-

tional to the frequency of population in the first age interval. This proportion is .107. The correspondence between *area* and *relative frequency* is a very useful property of the histogram.

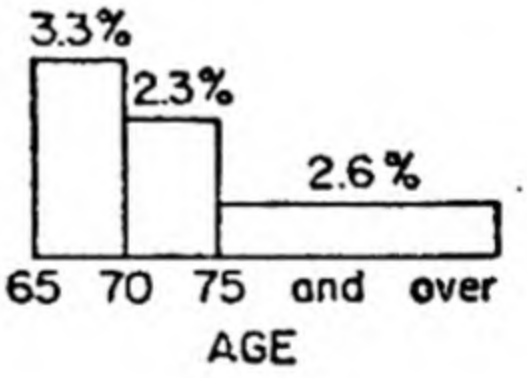
We have said that the relative frequency and the area of each rectangle are proportional. It is only when the rectangles have equal widths that the *height* is proportional to the frequency. If the 75-and-over open-end age class is made twice as wide on the horizontal scale as the previous classes, we would divide the relative frequency in this class by two (2.6% ÷ 2 = 1.3%) to determine the height. If the 75-and-over rectangle is made three times as wide, we would divide the relative frequency by three (2.6% ÷ 3 = .87%).[4] With intervals of unequal size, we can indicate the frequency *above* each rectangle rather than on a vertical frequency scale.

Age Distribution	*Relative Frequency*
65 to 69	3.3%
70 to 74	2.3%
75 and over	2.6%

Open-End Interval Twice as Wide as Previous Intervals



Open-End Interval Three Times as Wide as Previous Intervals



In addition to the histogram, another graphic presentation of the age distribution is the *frequency polygon* (Fig. 2.5). Using the same horizontal and vertical scale as for the histogram, we draw a frequency polygon by placing points directly above the midpoint of each class interval and across from the frequency in that interval, the adjacent points then being connected with straight lines. Note that, unlike the histogram, the area within any age interval of a frequency polygon does not represent the correct proportion of the total area

[4] The percentage of population 75-and-over could also be distributed unequally, e.g., 1.3% between 75 and 80 years; .9% between 80 and 85 years; .4% 85 years and over.

unless the frequency in that interval differs from the frequencies of adjacent intervals by a constant arithmetic amount.

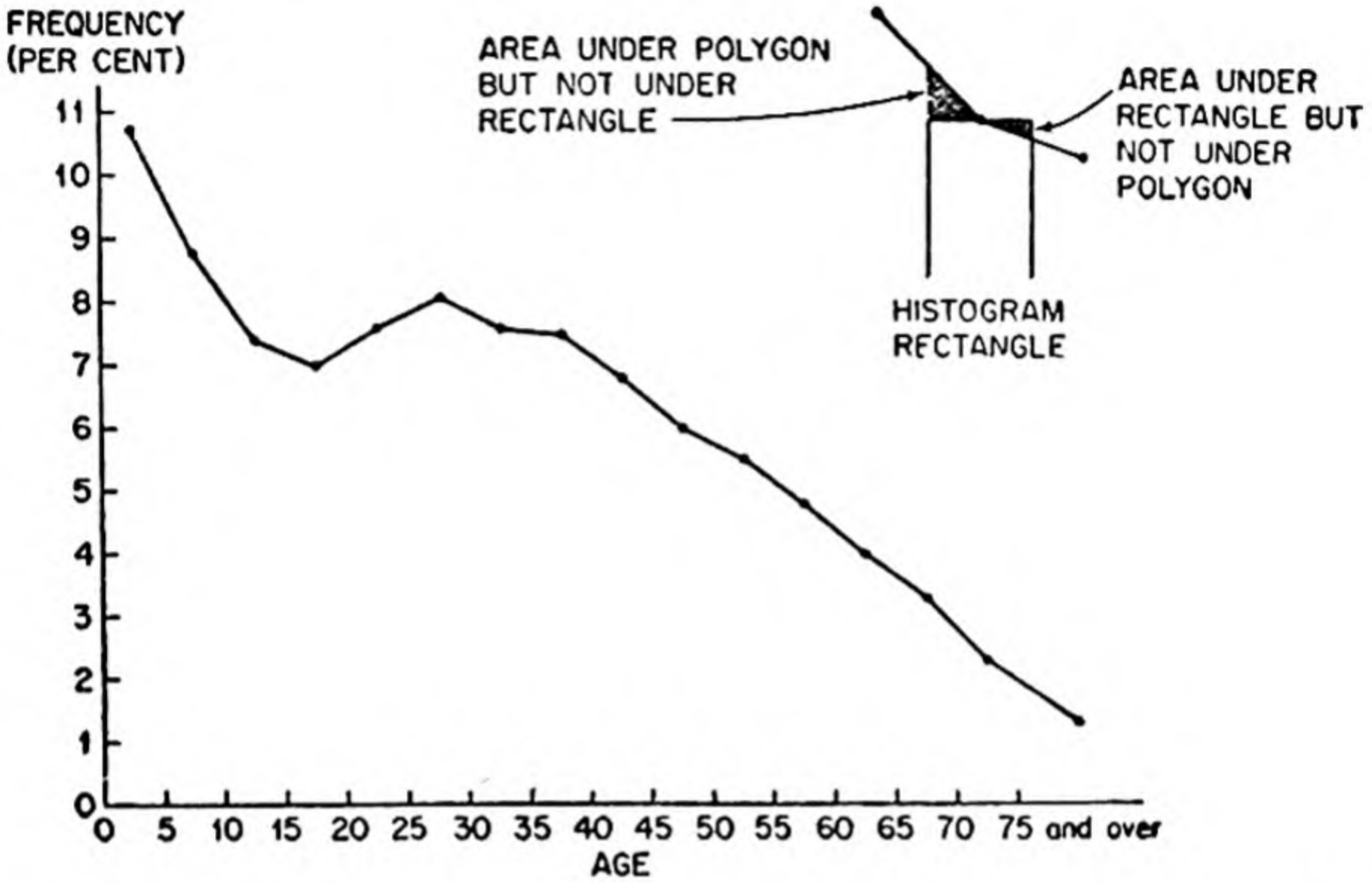


Fig. 2.5. Age distribution for the United States, 1950. (Frequency polygon for data of Table 2-3.)

The histogram and the frequency polygon give the number of cases for each class interval. We may want to know not only the frequency for any specified class interval (how many children are between 5 and 10 years) but also the cumulative frequency, i.e., how many children are under 10 years of age. The cumulative frequency columns of Table 2-5 indicate that about 29 million children in the United States are under ten years of age.

The *cumulative frequency polygon* of Fig. 2.6 gives the cumulative frequency distribution in graphic form. We indicate cumulative frequency under a given age with a point directly above the upper limit of the class interval and across from the cumulative frequency for that interval. The cumulative frequency under 10 years of age, for example, is indicated by a point directly above the upper limit of the class interval 5 to 9 years of age and across from the cumulative frequency of 19.5 per cent.

From Fig. 2.6 we can quickly tell what per cent of the population falls under a given age. If we draw a horizontal line from the 50th

Table 2-5. Age Distribution for the United States, 1950

Age	Frequency		Cumulative Frequency	
	Number (000's)	*Per cent*	*Number* (000's)	*Per cent*
(1)	(2)	(3)	(4)	(5)
Total	150,697	100.0		
Under 5 years	16,164	10.7	16,164	10.7
5 to 9	13,200	8.8	29,364	19.5
10 to 14	11,119	7.4	40,483	26.9
15 to 19	10,617	7.0	51,100	33.9
20 to 24	11,482	7.6	62,582	41.5
25 to 29	12,242	8.1	74,824	49.6
30 to 34	11,517	7.6	86,341	57.2
35 to 39	11,246	7.5	97,587	64.7
40 to 44	10,204	6.8	107,791	71.5
45 to 49	9,070	6.0	116,861	77.5
50 to 54	8,272	5.5	125,133	83.0
55 to 59	7,235	4.8	132,368	87.8
60 to 64	6,059	4.0	138,427	91.8
65 to 69	5,003	3.3	143,430	95.1
70 to 74	3,412	2.3	146,842	97.4
75 and over	3,855	2.6	150,697	100.0

Source: Bureau of the Census, U.S. Department of Commerce, *1950 Census of Population, Advance Reports*, Series PC-14, No. 5, October 31, 1952.

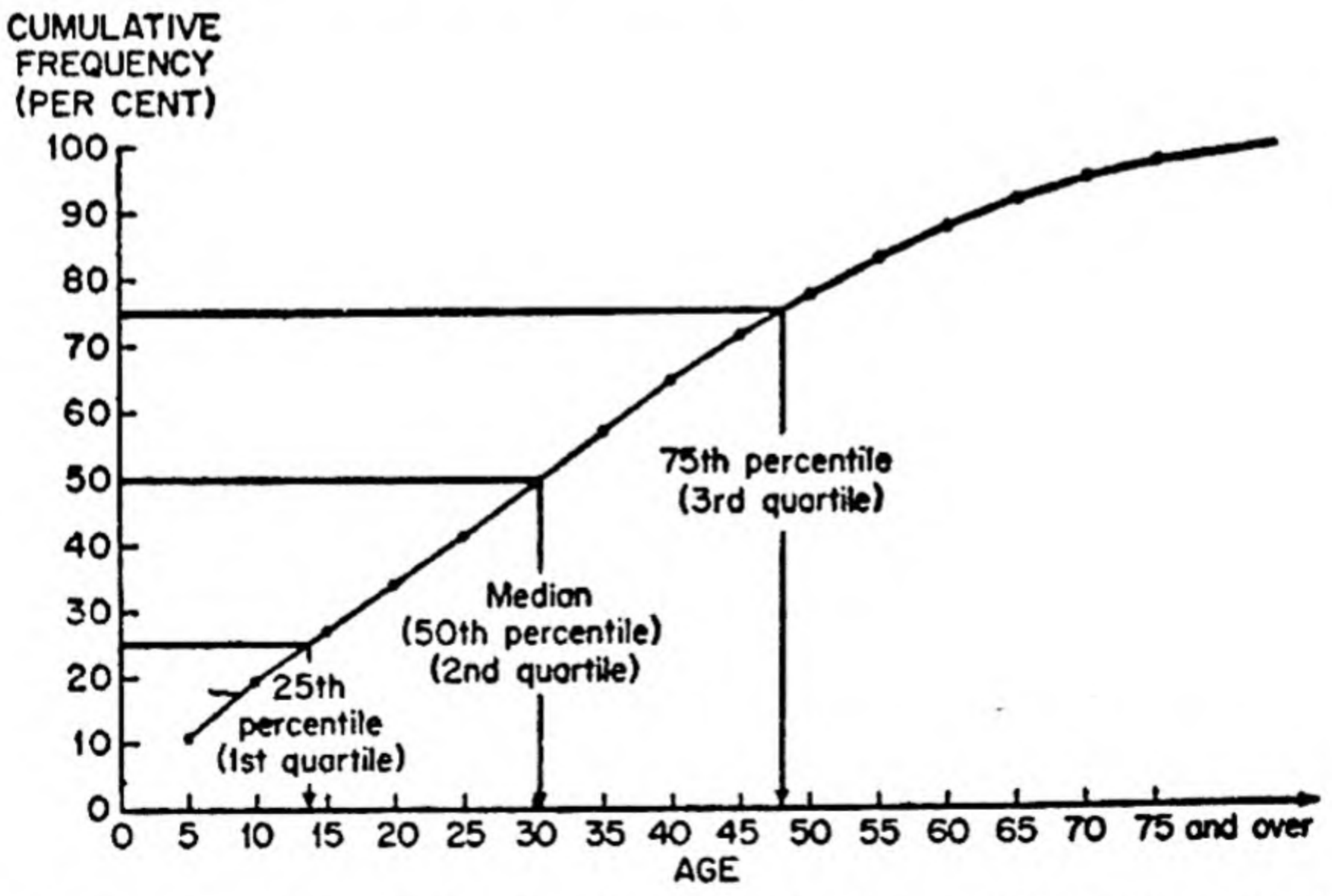


Fig. 2.6. Age distribution in the United States, 1950. (Cumulative polygon for columns 4 and 5 of Table 2-5.)

per cent of the vertical scale to the polygon and drop a line down to the horizontal scale, we find that 50 per cent of the population are less than about 30 years of age. The 30 is called the *50th percentile* or the *2nd quartile* or the *median*. On a histogram, 50 per cent of the area of the rectangles would be under age 30. Seventy-five per cent of the population are under about 48 years of age; the 48 is the 75th percentile (the 3rd quartile), and the 25th percentile (the 1st quartile) is about 14 years; 25 per cent of the population is younger than 14 years.

Symmetry and Skewness. The age distribution in the frequency polygon of Fig. 2.5 is a skewed and not symmetrical distribution. A skewed distribution lacks symmetry about a vertical axis through the center of the distribution. In an age distribution, there is a limit to how far one can go on the left (nobody is below age zero) but no such rigid limit is imposed on the right. The age distribution is said to be skewed to the right, in the direction of the tail, or positively skewed.

To illustrate skewness and symmetry, smoothed frequency polygons of three distributions are given in Fig. 2.7. The first, heights

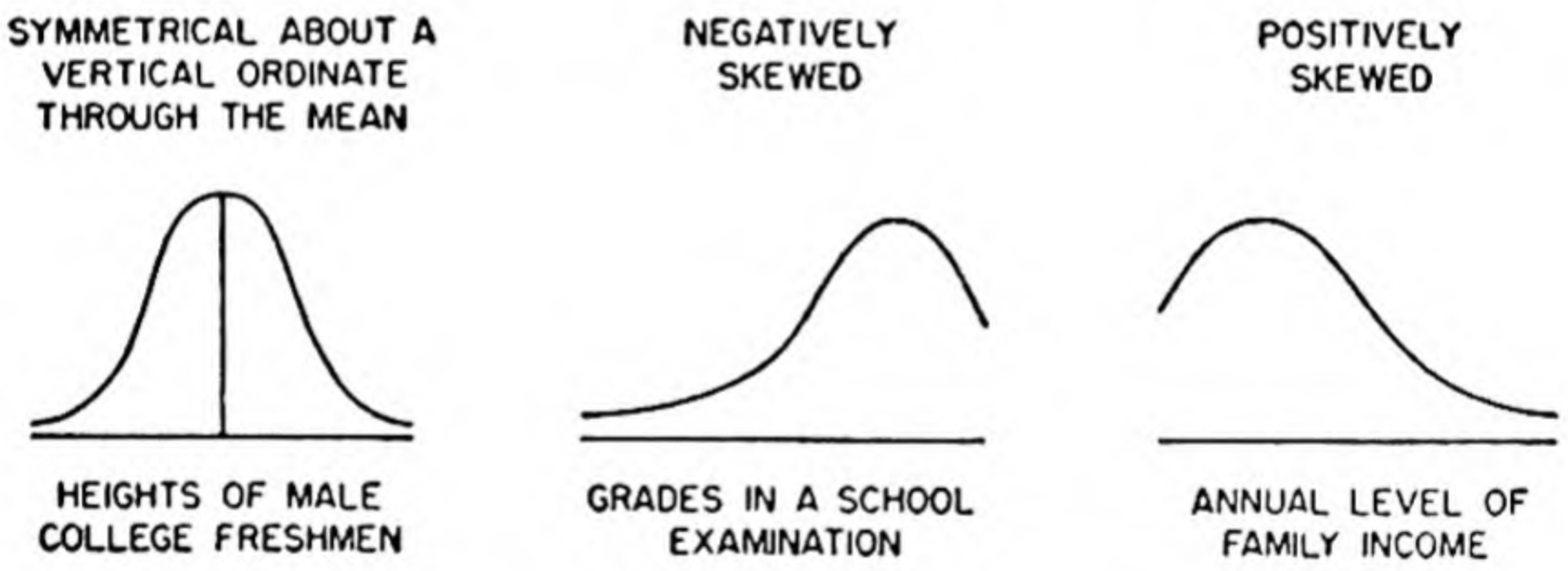


Fig. 2.7. Illustrations of symmetrical and skewed distributions.

of male college freshmen, is a symmetrical distribution. The second, grades in a school examination, is skewed to the left, or negatively skewed. Most of the scores on the school examination are between 70 and 90; the grades cannot go above 100, but can extend to the left as far as 0. The third polygon, family income distribution in the United States, is skewed to the right, or positively skewed. Income cannot go below zero, but can extend far to the right.

KEY TERMS

cumulative frequency distribution
cumulative frequency polygon
fiftieth percentile
first quartile
frequency polygon
histogram
median
pie chart
second quartile
seventy-fifth percentile
skewness
symmetry
third quartile
twenty-fifth percentile
vertical bar graph

REFERENCES

Dixon, W. J., and F. J. Massey, *Introduction to Statistical Analysis*, chap. 2. New York: McGraw-Hill Book Company, Inc., 1951.

Hagood, M. J., and D. O. Price, *Statistics for Sociologists*, chap. 5. New York: Henry Holt and Company, 1952.

Lindquist, E. F., *A First Course in Statistics*, chap. 4. Boston: Houghton Miffin Company, 1942.

Mode, E. B., *Elements of Statistics*, 2nd ed., chaps. 2 and 5. New York: Prentice-Hall, Inc., 1951.

Treloar, Alan E., *Biometric Analysis*, chap. 3. Minneapolis: Burgess Publishing Company, 1951.

EXERCISES

1. (a) Prepare three vertical bar graphs from the data of Table 2-2 showing the population by sex in (1) urban United States; (2) rural-nonfarm United States; and (3) rural-farm United States.

(b) Draw a fourth graph condensing the information of the first three by placing urban-rural residence on the horizontal scale and sex ratio on the vertical scale.

2. Given below are the scores on a 100-question introductory mathematics test taken by fifty students:

95	82	97	77	76	72	58	84	87	96	87	75	80
96	65	83	92	85	93	87	83	81	82	77	71	91
91	99	82	77	88	76	60	81	84	93	75	85	
90	78	86	70	88	76	76	80	86	94	96	84	

(a) Arrange these scores in order of size.

(b) Prepare a table with the following headings:

Interval limits	Interval midpoints	Frequency		Cumulative frequency	
		Number	Per cent	Number	Per cent

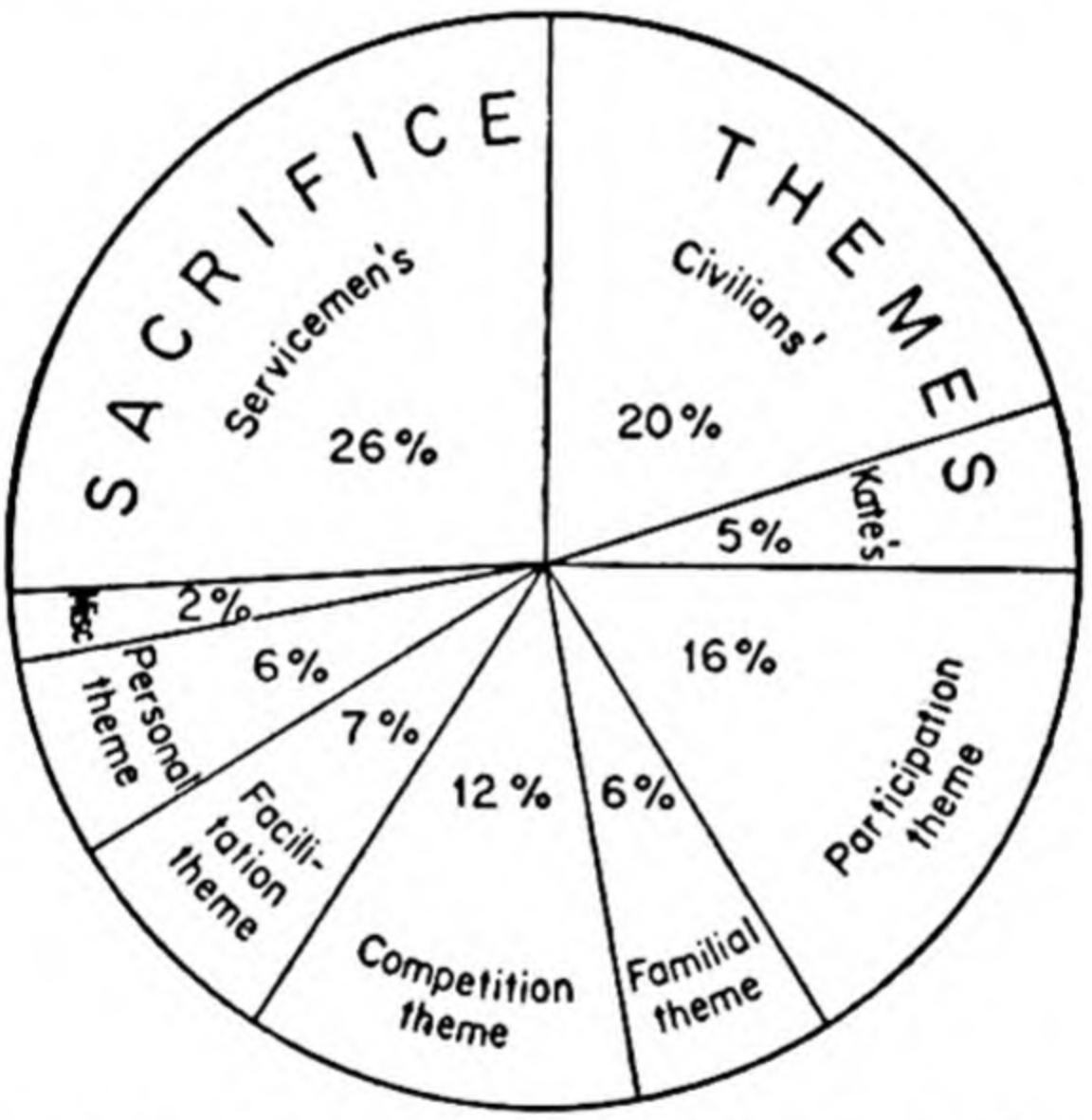
Divide the data into nine class intervals. Fill in the body of the data.

(c) Draw a histogram and a frequency polygon for this distribution.

(d) Draw a cumulative frequency polygon. Read from the cumulative polygon the approximate 50th percentile (the median), the 75th percentile, the 25th percentile.

3. The accompanying pie chart gives the proportion of time that Kate Smith spent on various themes in an 18-hour radio War Bond drive. To what theme did she devote about one-half of her time? Are there any striking omissions of themes?

What Smith Said: Time Distribution of Themes



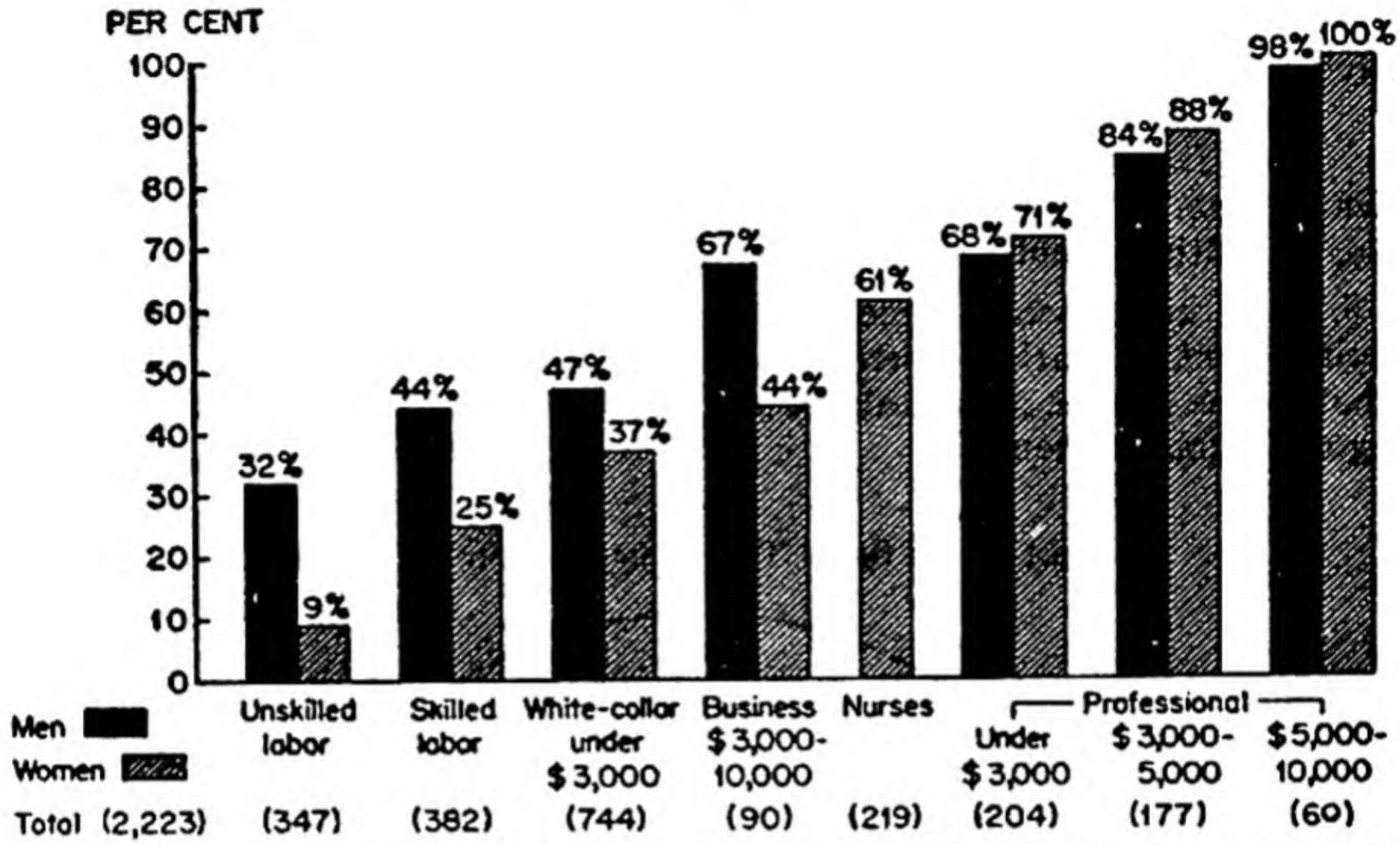
Source: Robert K. Merton, *Mass Persuasion* (New York: Harper and Bros., 1946), p. 50.

4. (a) Does the accompanying bar graph substantiate the description of Americans as a joining people? Is it the affiliated or unaffiliated persons who constitute a majority?

(b) What relationship appears to exist between participation in voluntary associations and socio-economic class?

(c) What are the class differences in the participation rates of women?

Percentage Belonging to One or More Associations by Class and Sex
(2,223 adult residents of New York City, 1934–35)

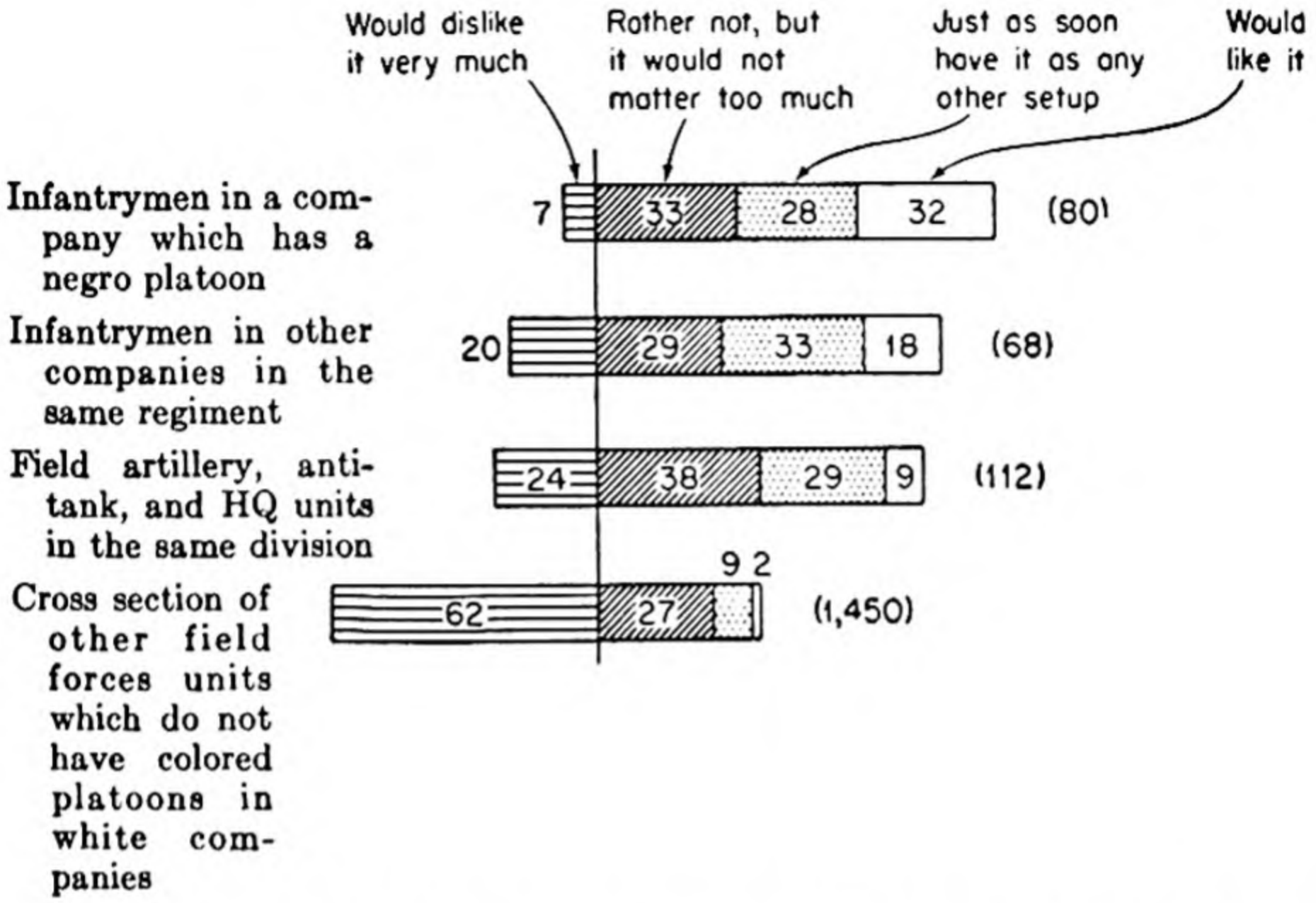


SOURCE: Mirra Komorovsky, "The Voluntary Associations of Urban Dwellers," *American Sociological Review*, *XI*, 686–698.

5. Does the graph indicate a relationship between *attitude toward* serving in a company containing Negro and White platoons and actually *being in* such a company? Can this be considered a "test case" in ordinary Negro-White relationships?

Attitudes toward Serving in a Company Containing Negro and White Platoons among Men Who Have Done So and Men Who Have Not
(Europe, June 1945)

Question: Some army divisions have companies which include Negro platoons and White platoons. How would you feel about it if your outfit was set up something like that?



SOURCE: Samuel Stouffer et al., *The American Soldier*, Vol. I (Princeton: Princeton University Press, 1949), p. 594. Data from ETO-82.

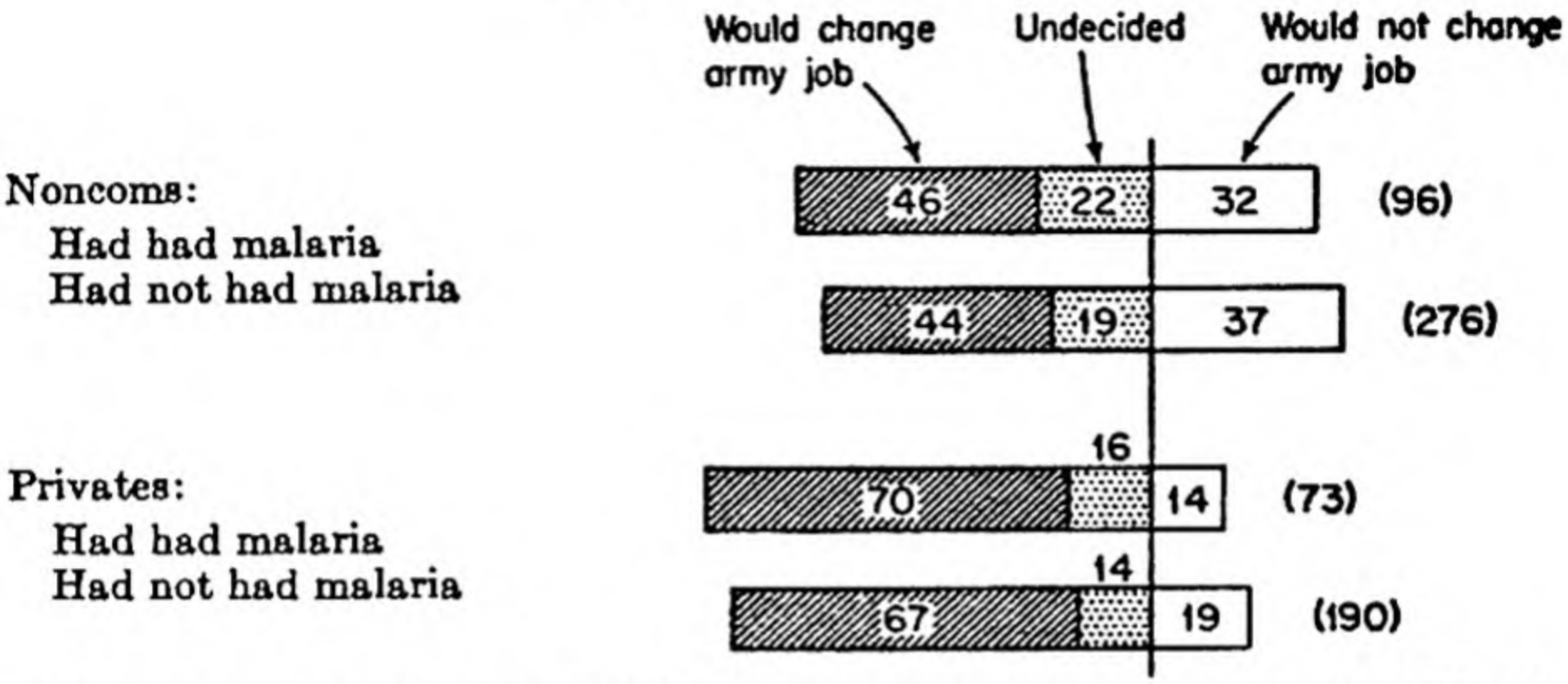
* The numbers following the bars are the number of cases on which percentages are based.

6. According to the graph, is (1) exposure to worse environmental conditions (as indicated by contracting malaria) or (2) being a private (and not a noncommissioned officer) a more significant variable in determining job attitudes?

Attitude toward Job among Soldiers in the Solomons by Whether or Not Respondents Had Had Malaria

(Noncombat troops in Guadalcanal, New Georgia, and Espiritu Santo who have been overseas 1 to 2 years, January 1944)

Percentage Distribution*



SOURCE: Samuel Stouffer et al., *The American Soldier*, Vol. I (Princeton: Princeton University Press, 1949), p. 348.

* The numbers following the bars are the number of cases on which percentages are based.

7. In *Patterns for Industrial Peace*, Whyte hypothesizes that attitudes can be changed by changing patterns of interaction. He gives a diagrammatic representation, reproduced on page 33, of the changing patterns of interaction between management and union in a steel plant at four company levels and four corresponding union levels. The *direction* of the arrows indicates the origination of action and their *relative thickness*, the frequency of such action.

(a) When are the channels of upward communication to the top of the management structure first open? When are they greatest?

(b) During what period do the frequency of upward and downward communications within the management and within the union structure most closely approximate each other?

(c) Does top management always observe rigorously the channels of the organization? Does it ever do any by-passing?

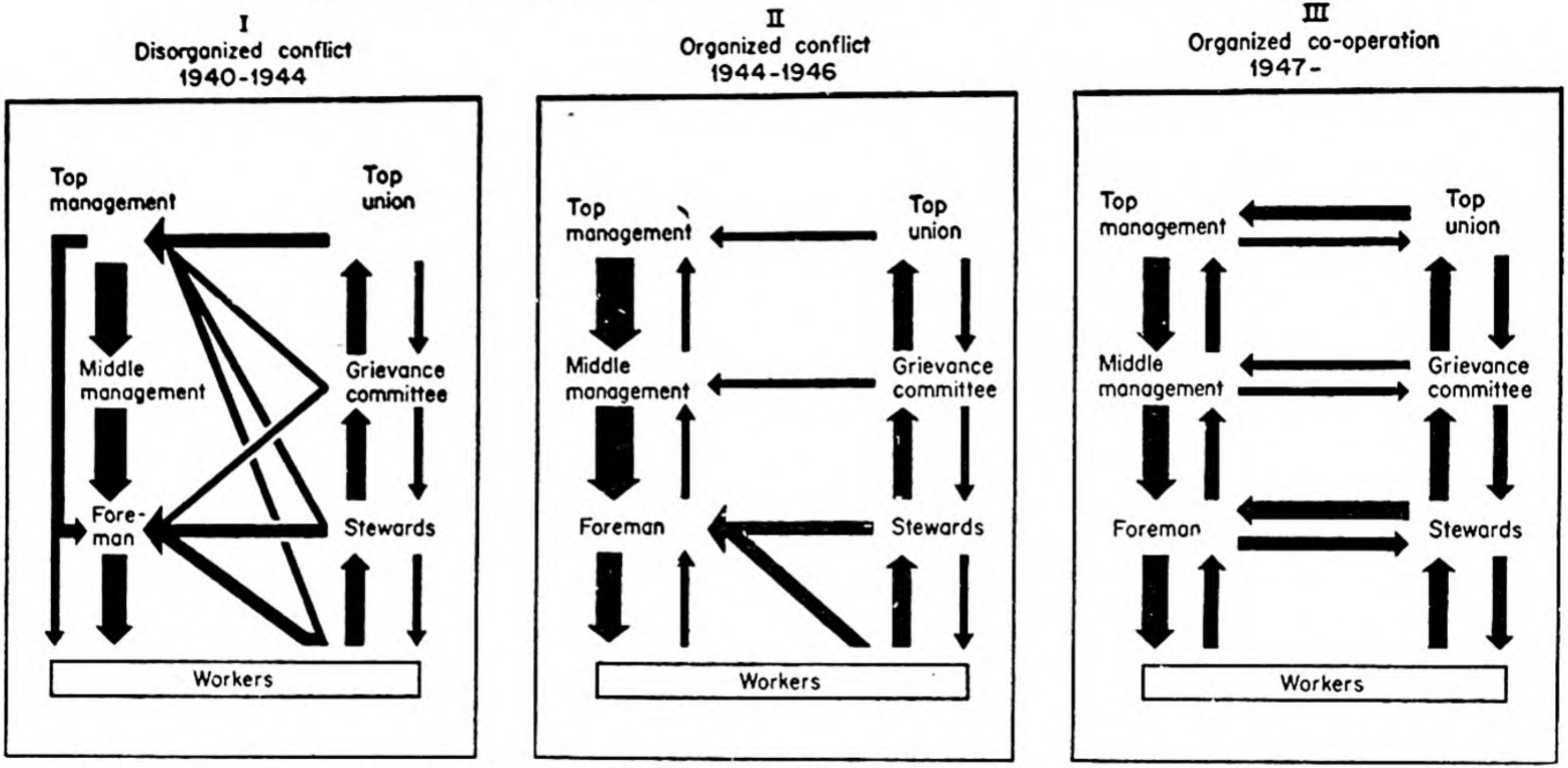
(d) What part of the management structure is under the greatest pressure from all sides during the first period (aside from the mass-demonstration pressure on top management)? Does this pressure persist during the three periods?

(e) How do the union's originations on management differ from one period to another (in direction and frequency)?

(f) During what period does management utilize two channels to get things done—the union structure as well as the management structure?

(g) Could we assume from the diagram that changing *attitudes* have changed the patterns of interaction rather than vice versa?

Patterns of Interaction between Management and Union of an Industrial Plant



Source: William F. Whyte, *Patterns for Industrial Peace* (New York: Harper and Bros., 1951), p. 165.

CHAPTER 3

MEASURES OF CENTRAL VALUE AND DISPERSION

3.1. Measures of Central Value

The monthly rent for five dwelling units is \$40, \$45, \$50, \$60, and \$80, a rental variation from \$40 to \$80. If we want to use only one figure to represent the central value, the average rent of these five dwelling units, what figure will it be?

The Arithmetic Mean. One measure of central value is the arithmetic mean. The mean is the center of gravity of the distribution. If the frequency distribution were plotted on heavy cardboard, and then cut out with a scissors, the point at which it could be balanced would be its mean. It is equal to total value divided by the number of cases—in the example above, the total monthly rent bill for all five units divided by five. Its formula is $\Sigma X/n$. The formula ΣX, read sum of X, is the total value of the variable rent, and n is the number of dwelling units. The mean monthly rent for these five dwelling units is:

$$\bar{X} = \frac{\Sigma X}{n} = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{n} \qquad (1)$$

$$= \frac{\$40 + \$45 + \$50 + \$60 + \$80}{5}$$

$$= \frac{\$275}{5} = \$55$$

X is the symbol for the variable, monthly rent, and $\bar{X}$, read X bar, is the mean of the observed cases, that is, the mean monthly rent for the five dwelling units. Note that the arithmetic mean of \$55 does not equal any actual rental. Note also that the sum of the deviations from the mean equals zero, the sum of the deviations below the mean being exactly equal to the sum of the deviations above the mean.

$$\Sigma(X - \bar{X}) = (\$40 - \$55) + (\$45 - \$55) + (\$50 - \$55) + (\$60 - \$55) + (\$80 - \$55)$$
$$= \quad -15 \qquad -10 \qquad -5 \qquad +5 \qquad +25$$
$$= 0$$

If we want to compute the mean monthly rent for one hundred dwelling units instead of five, the use of the formula $\Sigma X/n$, which sums the monthly rent of all dwelling units, becomes burdensome, especially if no computing machine is available. (See Table 3-1.)

Table 3-1. Monthly Rent and Computation of the Mean for One Hundred Dwelling Units
(Ungrouped data)

$25	$40	$50	$55	$60	$67	$72.50	$80	$90	$102
26	40	50	55	62	67.50	72.50	83.50	90	102
30	42	50	59	62	70	74	83.50	93	105
30	42.50	50	60	62.50	70	74	84	93	107.50
30	42.50	50	60	62.50	70	77.50	84	94	110
33	45	50	60	64	70	77.50	84	95	110
35	45	52.50	60	64	70	78	85	97.50	112.50
38	47.50	52.50	60	64	70	80	87.50	100	117.50
40	47.50	54	60	65	70	80	87.50	100	120
40	47.50	55	60	65	72.50	80	90	100	122.50

$$\bar{X} = \frac{\Sigma X}{n} = \frac{\$6874}{100} = \$68.74$$

The computation of the arithmetic mean can be simplified by grouping the data into class intervals. The monthly rentals of the one hundred dwelling units extend from $25 to $122.50, a $97.50 range. Rentals appear to cluster at $30, $40, $50, etc., so these points will be used as the midpoints of the intervals in order to minimize the errors which arise from grouping.

If the range of approximately $100 is divided into intervals of $10 each, there will be ten intervals, the first extending from $25 to $34.99 with a midpoint of $30, the tenth from $115 to $124.99 with a midpoint of $120. The monthly rental distribution is a continuous one.

To compute the mean from grouped data, the number of dwelling units in each interval (column 3 of Table 3-2) is multiplied by the midpoint of the interval (column 2) and summed, which gives a total rent bill of $6,880 for the one hundred dwelling units (column 4). The mean monthly rent is $6,880 divided by 100 (formula 2), or $68.80 (as compared with a mean of $68.74 from the ungrouped data of Table 3-1). The mean computed from the grouped data is not precisely accurate because of the assumption that the midpoint re-

Table 3-2. Monthly Rent Distribution and Computation of Mean for One Hundred Dwelling Units
(Grouped data)

(1)	(2)	(3)	(4)	(5)	(6)
Monthly Rent (*X*)	*Interval Midpoint* (*m.p.*)	*Frequency* (*f*)	*Interval Total* $f \times m.p.$	*Deviation from Guessed Mean (in interval units)* (*d'*)	*Frequency Times Interval-Deviation from Guessed Mean* (*fd'*)
$ 25– 34	$ 30	6	$ 180	−4	−24
35– 44	40	9	360	−3	−27
45– 54	50	14	700	−2	−28
55– 64	60	19	1,140	−1	−19
65– 74	70	16	1,120	0	0
75– 84	80	12	960	1	12
85– 94	90	9	810	2	18
95–104	100	7	700	3	21
105–114	110	5	550	4	20
115–124	120	3	360	5	15
Sum:		100 (= n)	6,880		−12

Computation of Mean:

First Method:
$$\bar{X} = \frac{\Sigma[(f) \times (m.p.)]}{n} \tag{2}$$
$$= \frac{\$6880}{100}$$
$$= \$68.80$$

Second Method:
$$\bar{X} = \text{Guessed mean} + \left(\frac{\Sigma fd'}{n}\right) C \tag{3}$$
$$= \$70 + \left(\frac{-12}{100}\right) 10$$
$$= \$70 + (-.12)10$$
$$= \$70 - \$1.20 = \$68.80$$

where C equals the size of the interval.

presents the arithmetic mean of the interval, but it is usually a sufficiently good approximation if the interval is not too large and the distribution is symmetrical, or only moderately asymmetrical.

The mean of a frequency distribution can also be computed by a guessed mean method, which reduces arithmetic computations by utilizing interval deviations from a guessed mean. (See formula 3.)

The midpoint of the interval estimated to contain the mean is chosen as the guessed mean; the answer will be the same whatever the choice. In Table 3-2, the mean is guessed at $70. The $70 midpoint is given a 0 value, this value being zero deviations away from the guessed mean. One interval below the guessed mean ($70 *minus* $10, or $60) is given a -1 value; two intervals below ($70 *minus* $20, or $50) is valued at -2. One interval above the guessed mean ($70 plus $10, or $80) is considered $+1$, etc. (column 5 of Table 3-2).

These interval deviations from the guessed mean (d') are multiplied by the corresponding frequency in each interval, summed, and divided by n to give an average interval deviation from the guessed mean [$\Sigma fd'/n = -.12$ (column 6 of Table 3-2)]. The average interval deviation from the guessed mean may be positive or negative, according to whether the guessed mean is above or below the true mean. To convert the average *interval* deviation from the guessed mean into an average *dollar* deviation, we multiply by the size of the interval ($10). The average dollar deviation is $-.12 \times \$10$, or $-\$1.20$.

The sum of the deviations from the true mean always equals zero. (If the guessed mean of our problem were actually the true mean, the average deviation would equal zero.) Adding the average dollar-deviation ($-\$1.20$) to the guessed mean gives the true mean. [$\$70 + (-\$1.20) = \$68.80$.]

In a symmetrical distribution, the mean is at the center of symmetry. The monthly rental distribution of Table 3-2 is not symmetrical, but skewed somewhat to the right. The frequency polygon is given in Fig. 3.1.

If a frequency distribution has an open-end interval, an arbitrary decision is made about the midpoint of this open-end interval in order to compute a mean. This decision is based wherever possible on knowledge of the data. If it is desired, for example, to compute the midpoint of an open-end monthly rental interval of $100-and-over for tenant-occupied dwelling units in the city of Chicago, 1950, the decision is based upon what is known about the distribution of rents at the $100-and-over level in Chicago.

The Median. The median is the middle value of the variable when the values are arranged according to size. If there is no middle value, the median is considered to be the interpolated middle value. One-half the cases have higher values than the median, the other half,

lower values. In a histogram, a vertical line through the median would divide the total area into two equal parts.

If five dwelling units have monthly rentals of $40, $45, $50, $60, and $80, the median rental is $50, the middle value. There are five cases, and the middle case is $\frac{1}{2}(n + 1) = \frac{1}{2}(5 + 1) = 3$.

If six dwelling units (an even number of units) have monthly rentals of $40, $45, $50, $60, $70, and $80, the median is regarded as the average of the two middle values, which is $55 [($50 + $60)/2.] The $55 is not really the middle value; there is no middle value with an even number of cases.

The median rental for seven dwelling units with rentals of $40, $50, $50, $50, $60, $70, and $80 is the fourth value, or $50. Note that only one rental is below $50, three are above. The meaningfulness of the $50 median is questionable.

With continuous data grouped into a frequency distribution, the median is the interpolated middle value, the value for the $\frac{1}{2}$ *n*th case. To find the median, we first determine what interval contains the $\frac{1}{2}$ *n*th case, and then interpolate. The assumption is made, in interpolating, that the cases are evenly distributed throughout the interval.

$$\text{Median} = \begin{array}{l}\text{Lowest point}\\ \text{in interval}\\ \text{containing}\\ \frac{1}{2}\,n\text{th case}\end{array} + \left[\frac{\begin{array}{l}\text{Frequency needed}\\ \text{in median-interval}\\ \text{to get to } \frac{1}{2}\,n\end{array}}{\begin{array}{l}\text{Frequency in}\\ \text{median interval}\end{array}}\right] \times \begin{array}{l}\text{Size of}\\ \text{interval}\end{array} \tag{4}$$

In the 100 dwelling-unit problem (Table 3-2), $\frac{1}{2}$ *n* is the fiftieth case. To determine which interval contains the fiftieth case, we add up the cases until we come to the fiftieth: 6 + 9 + 14 + 19 equals 48. The interval containing the fiftieth case is $65 − $74; then $65 is the lowest point in the interval. Two cases, out of the 16 cases in the interval, are needed in the interval to reach the fiftieth case. The size of the interval is $10. Therefore

$$\text{Median} = \$65 + \left(\frac{2}{16}\right) \times \$10 = \$66.25$$

To check the computation, we can start at the top of the distribution and *subtract* from the highest point in the interval

$$\text{Median} = \$75 - \left(\frac{14}{16}\right) \times \$10 = \$66.25$$

Note the minus sign in the latter equation. We want to stay in the interval containing the $\frac{1}{2}$ nth case. Hence we add to the lowest point in the interval, but subtract from its highest point.

In the calculation of the median it is assumed that the cases in the interval in which the median lies are uniformly distributed throughout the interval. This assumption is usually sufficiently accurate if the class interval is not too large and the distribution not highly skewed.

The Mode. The mode is the most typical, the most frequent value. Of the several ways of defining the mode of a frequency distribution, the simplest definition is the midpoint of the interval of greatest frequency. By this definition, the mode for the 100 rental units of Table 3-2 is \$60.

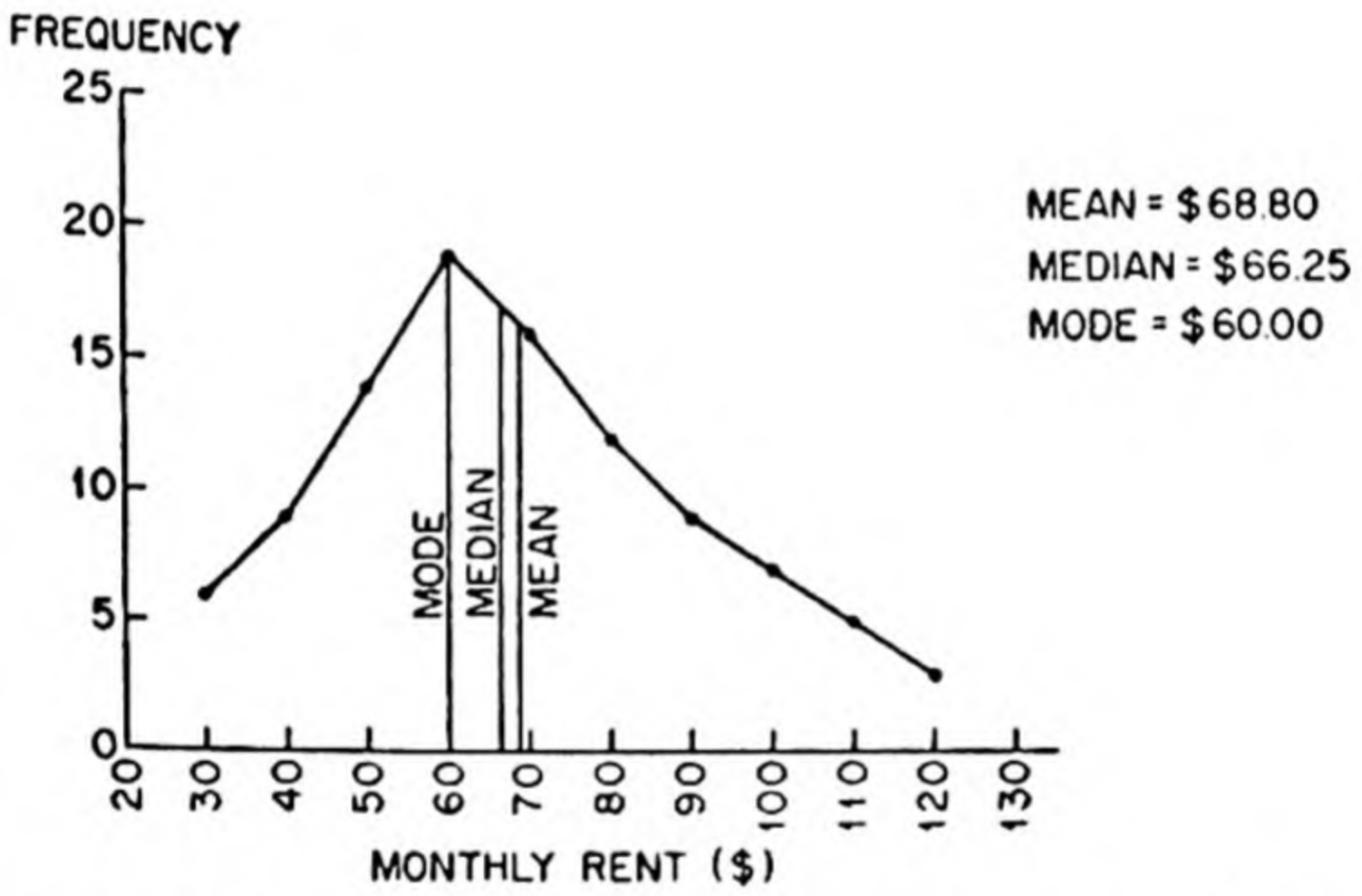


Fig. 3.1. Monthly rent for one hundred dwelling units. (Frequency polygon for data of Table 3-2.)

The frequency polygon for the 100 dwelling unit monthly-rental distribution is drawn in Fig. 3.1, and the mean, median, and mode are superimposed upon it.

In a symmetrical distribution, the mean, median, and mode coincide at the center of symmetry. In a moderately skewed distribution, the mean is the measure most influenced by the skewness, and the median lies about two-thirds of the distance from the mode to the mean. The mean is most influenced by extreme cases, since all values of the variable are used in its calculation.

Two other measures of central value, not defined here, are the geometric and harmonic mean.

Which Measure of Central Value to Use? The term "measure of central value" seems to imply some concentration of values near the center of the distribution. If there is no such concentration, the use of a measure of central value alone is questionable.

In a fairly symmetric distribution, the mean, median, and mode are not far apart, so that the choice of a measure will not materially affect the result.

The arithmetic mean has useful mathematical properties. We have mentioned one: the sum of the deviations from the mean is zero. Another is the relative stability of the mean in repeated samples from the same universe in some common types of distributions. A third is the additive property of the arithmetic mean. If, for example, all ten classes in a school take an attitude test on radicalism-conservatism, and the mean of each class is weighted by the respective number of students, then the sum of the weighted means divided by the number of students will give an average for the entire school. The arithmetic mean can be manipulated mathematically more than the other averages.

Highly skewed distributions and distributions with open-end intervals may make the median or mode a more representative measure than the mean. The median is practical in cases where information is desired on the relative position of a person or thing—whether, for example, a person is in the upper or lower half among participants in a State Board examination.

The mode is feasible when one value or class interval predominates. If, in a Maine town, 70 per cent of the registered voters are Republican, we can say that the typical or modal voter is Republican. The mode is a practical measure for qualitative data. If 90 per cent of the population say they belong to the middle class, then the middle class represents the modal or most typical class according to the respondents' statements. Where the percentage is as high as 90, it is probably more meaningful to indicate the percentage than to use any kind of average.

The measure of central value selected is often determined by the use of the data. What average should be used for income tax payments? The Federal Government, concerned with total Federal income tax revenues, may compute the mean tax bill as an average,

since the mean tax bill, when multiplied by the number of taxpayers, gives total revenue. Labor unions, on the other hand, concerned with what the average or typical taxpayer pays, will probably compute the median or modal tax payment.

The average selected is sometimes determined by the desired effect. In the case of a lottery, the ticket seller, anxious to sell tickets, may advertise the average lottery winning of the previous year as perhaps $500; he may even publicize the chances of winning $25,000 for only 25¢. The $500 average, which is the arithmetic mean of the winnings, is influenced by a few extreme cases, the ones who win $25,000 and $10,000. The distribution of the winnings is highly skewed to the right. The most typical winning, the mode, may be only $1. The losers may not be included at all in the determination of the mean winnings.

A manufacturer, anxious to attract workmen to his plant, announces that the average take-home pay of workers in his plant is $100 a week. This arithmetic mean of $100 may have been strongly influenced by the wages of several highly skilled workmen. The median and mode are probably lower. A university may publish the mean annual income of its twenty-five-year alumni class, using the average most influenced by extreme cases. The few extremely high incomes may have resulted not from the acquisition of learning at the university, but from the inheritance of family fortunes.

KEY TERMS

arithmetic mean	**interpolation**	**mode**
grouped data	**median**	**ungrouped data**

REFERENCES

Dixon, W. J., and F. J. Massey, *Introduction to Statistical Analysis*, chap. 3. New York: McGraw-Hill Book Company, 1951.

Freund, J. E., *Modern Elementary Statistics*, chap. 4. New York: Prentice-Hall, Inc., 1952.

McNemar, Quinn, *Psychological Statistics*, chap. 3. New York: John Wiley & Sons Inc., 1951.

Mode, Elmer, *Elements of Statistics*, 2nd ed., chap. 3. New York: Prentice-Hall, Inc., 1951.

Treloar, Alan E., *Biometric Analysis*, chap. 4. Minneapolis: Burgess Publishing Company, 1951.

EXERCISES

1. Included on a questionnaire survey of rooming-house residents was the question: How many times a week do you read a newspaper? The replies of a sample of one hundred rooming-house residents are given below:

Number of Times a Week Newspaper Read	*Frequency of Replies*
0– 1	16
2– 3	17
4– 5	24
6– 7	15
8– 9	12
10–11	10
12–13	6

(a) Is this variable discrete or continuous? What are the limits and midpoint of the first interval?

(b) What is the mean number of times the newspaper is read among this rooming-house sample? (Compute the mean by two different methods; one method should serve as a check on the other.)

(c) Draw a frequency polygon.

(d) Determine the cumulative frequency distribution for this sample and draw a cumulative frequency polygon.

(e) Does the answer to this question show whether or not we have a well-read sample of rooming-house residents? Why or why not? What additional questions might be asked to get at this information?

2. One hundred people are interviewed in a town by a public opinion polling agency. The frequency distribution gives the ages of the people interviewed. (Age is given as of the last birthday.)

Ages of People Interviewed	*Frequency*
80–89	1
70–79	1
60–69	3
50–59	10
40–49	28
30–39	20
20–29	21
10–19	16
	$n = 100$

(a) What is the mean age? The median? The mode?

(b) Draw a histogram and insert the three measures of central value.

(c) Which measure of central value is dependent upon the value of all the cases in the distribution? Which is dependent not upon the value of all the cases, but upon the position of all the cases?

3. The accompanying table gives the percentage distribution of families in the United States according to their 1949 income level.

Income of Families in the United States, 1949*

1949 Money Income Level	*Number* (000's)	*Per Cent*
Total reporting	36,441	100.0
Under $500	3,129	8.6
$ 500– 999	2,496	6.8
1,000–1,499	2,664	7.3
1,500–1,999	2,717	7.5
2,000–2,499	3,378	9.3
2,500–2,999	3,292	9.0
3,000–3,499	3,989	10.9
3,500–3,999	3,179	8.7
4,000–4,499	2,620	7.2
4,500–4,999	1,786	4.9
5,000–5,999	2,864	7.9
6,000–6,999	1,557	4.3
7,000–9,999	1,714	4.7
10,000 and over	1,054	2.9

SOURCE: *1950 Census of Population, Preliminary Reports*, PC-7, No. 2, April 11, 1951.

* The difference between the sum of the figures and the total is due to rounding.

(a) Compute the mean and median. Use $300 as the midpoint for incomes under $500, $12,000 as the midpoint for incomes of $10,000 and over.

(b) Draw a histogram and insert the mean and the median. Why would you expect the mean to be higher than the median?

(c) Which measure, the mean or the median, better indicates the standard of living in the United States? If it were found that the mean income of Spain is higher than that of England, could you assume a higher standard of living in Spain than in England?

(d) Which measure would you use if you wanted to add the average income in the United States to the averages in Canada and Mexico to give a weighted average income for North America?

(e) If *A* knows the mean income, *B* knows the median income, and *C* knows both the mean and the median, what characteristic of the distribution does *C* alone know?

4. Tell which measure of central value applies in the three instances given below:

(a) A case picked at random is as likely to be above as below this measure.

(b) A case picked at random is most likely to fall in the interval containing this measure.

(c) It is possible that no case in the distribution has exactly this value, even though the value is typical of the whole distribution.

5. At a certain college there are five hundred freshmen, three hundred sophomores, and two hundred upper-classmen. The mean age for freshmen is 18; for sophomores, 20; and for upper-classmen, 22. What is the mean undergraduate age?

If you know the median age of students in each class instead of the mean age, can you compute an average undergraduate age? Why or why not?

6. Can a distribution have two means? Two medians? Two modes?

3.2. Measures of Dispersion

Distributions may have the same arithmetic mean, and yet differ in their scatter. In the very simple example given in Table 3-3,

Table 3-3. Computation of Mean and Measures of Dispersion for the Age Distribution of Two Hypothetical Families

	Family A				Family B			
Family Members	*Age*	*Mean Age*	*Deviations from Mean*	*Squares of Deviations from Mean*	*Age*	*Mean Age*	*Deviations from Mean*	*Squares of Deviations from Mean*
	(X)	$(\bar{X})$	$(x = X - \bar{X})$	$(x)^2$	(X)	$(\bar{X})$	$(x = X - \bar{X})$	$(x)^2$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Father	60	25	35	1,225	39	25	14	196
Mother	45	25	20	400	38	25	13	169
Siblings	10	25	−15	225	18	25	−7	49
	8	25	−17	289	16	25	−9	81
	2	25	−23	529	14	25	−11	121
Sum	125		0	2,668	125		0	616
Mean	25				25			

Measures of Dispersion:

Range: Highest value minus lowest value (5)

$60 - 2 = 58$ $\quad$ $39 - 14 = 25$

Mean Deviation: $\dfrac{\Sigma|x|}{n}$ $\quad$ $\dfrac{\Sigma|x|}{n}$ (6)

$= \dfrac{110}{5} = 22$ $\quad$ $= \dfrac{54}{5} = 10.8$

Variance: $\dfrac{\Sigma(x^2)}{n}$ $\quad$ $\dfrac{\Sigma(x^2)}{n}$ (7)

$= \dfrac{2668}{5} = 534$ $\quad$ $= \dfrac{616}{5} = 123$

Standard Deviation: $\sqrt{\dfrac{\Sigma(x^2)}{n}}$ $\quad$ $\sqrt{\dfrac{\Sigma(x^2)}{n}}$ (8)

$= \sqrt{534} = 23.1$ $\quad$ $= \sqrt{123} = 11.1$

two families have the same mean age of 25, but the variation in age between the two families differs considerably.

The range is probably the simplest measure of dispersion. It is the difference between the highest and the lowest value observed. The range of ages in Family *A* is 58 years (the difference between 60 and 2); the Family *B* range is 25 years (the difference between 39 and 14). The range is based on the two extreme observations and takes no account of the rest of the distribution.

Measures of dispersion *around the mean* have fruitful statistical application. One such measure of dispersion might conceivably be the average of the deviations from the mean. But the sum of the deviations from the mean equals zero (column 4 of Table 3-3); hence the average deviation from the mean must equal zero. Note that x stands for deviation from the mean, or $X - \bar{X}$. Deviations from the mean are represented symbolically by lower case letters.

If we disregard the plus and minus signs in computing the average deviation from the mean, we get a measure of dispersion known as the mean deviation. It is defined as $\Sigma|x|/n$, where $|x|$ stands for the absolute value of the deviations from the mean. (The absolute value of a number is equivalent to the number without its positive or negative sign.) For Family *A*, the absolute value of the deviations from the mean age of 25 are $35 + 20 + 15 + 17 + 23$, totaling 110 (column 4 of Table 3-3). The total of 110 divided by five (the number of family members) gives a mean deviation of 22. For Family *B* the mean deviation is 10.8. The mean deviation is rarely used as a measure of central value in sociological studies.

An alternative measure of dispersion around the mean can be gotten by *squaring* each deviation from the mean (column 5, Table 3-3), adding, and computing an average of the squared deviations.

$$\frac{\Sigma(X - \bar{X})^2}{n} \quad \text{or} \quad \frac{\Sigma(x^2)}{n}$$

$$= \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + (X_3 - \bar{X})^2 + (X_4 - \bar{X})^2 + (X_5 - \bar{X})^2}{n}$$

This measure, called the variance, will be interpreted in later statistical theory. For Family *A*, the squared deviations from the mean are $1{,}225 + 400 + 225 + 289 + 529$, which adds to 2,668; dividing 2,668 by 5 gives a variance of 534. The variance is symbolized as σ^2 if it is the variance of a universe, by s^2 if it is the variance of a

sample. (The terms "universe" and "sample" will be defined in Chapter 5.)

To partially nullify the effects of squaring the deviations from the mean, we can extract the square root.

$$\sqrt{\frac{\Sigma(X-\bar{X})^2}{n}} \quad \text{or} \quad \sqrt{\frac{\Sigma(x^2)}{n}}$$

$$= \sqrt{\frac{(X_1-\bar{X})^2+(X_2-\bar{X})^2+(X_3-\bar{X})^2+(X_4-\bar{X})^2+(X_5-\bar{X})^2}{n}}$$

This positive square root of the variance is called the standard deviation. It is symbolized by σ if it is the standard deviation of a universe and by s if it is the standard deviation of a sample. For Family *A*, the standard deviation is about 23; for Family *B*, it is about 11.

Note the influence of extreme cases on the variance and on the standard deviation. In the variance, all the deviations from the mean are squared. Squaring extreme cases far removed from the mean magnifies the influence of these cases. This influence is only partly removed by extracting the square root in the standard deviation. *Each* deviation is squared to compute the variance, but the square root is extracted only from the *average* of the squared deviations to compute the standard deviation. An increase of one unit in deviation is more important if it is a considerable distance from the mean than if it is very close to the mean.

Table 3-4 gives the computation of the mean and the standard deviation for the age distribution of sixty members of a family clan. When the standard deviation is computed from grouped data, the formula $\sqrt{\Sigma(x^2)/n}$ is modified to take into account the frequencies in the different intervals. The standard deviation formula used in Table 3-4 involves the computation of deviations from a guessed mean, with a correction factor compensating for the use of a guessed rather than the true mean.[1]

[1] Formulas for computing the standard deviation using the observational values, without getting deviations from a mean:

For Ungrouped Data:

$$\sigma = \sqrt{\frac{\Sigma(X^2)}{n} - \left(\frac{\Sigma X}{n}\right)^2} \quad \text{or} \quad \sigma = \frac{1}{n}\sqrt{n\Sigma X^2 - (\Sigma X)^2}$$

$$\sigma = C\sqrt{\frac{\Sigma f(d'^2)}{n} - \left(\frac{\Sigma fd'}{n}\right)^2} \tag{9}$$

where C equals the size of the class interval, and d' equals the deviation from the guessed mean in interval units.

Table 3-4. Computation of the Mean and Standard Deviation for the Age Distribution of a Hypothetical Family Clan

Age	*Interval Midpoint*	*Frequency* (f)	*Interval Deviation from Guessed Mean* (d')	fd' (col. 3 times col. 4)	$f(d'^2)$ (col. 4 times col. 5)
(1)	(2)	(3)	(4)	(5)	(6)
15–19	17.5	2	−3	−6	18
20–24	22.5	7	−2	−14	28
25–29	27.5	12	−1	−12	12
30–34	32.5	19	0	0	0
35–39	37.5	13	1	13	13
40–44	42.5	6	2	12	24
45–49	47.5	1	3	3	9
Sum		60 (= n)		−4	104

Mean:

$$\bar{X} = \text{Guessed Mean} + \left(\frac{\Sigma fd'}{n}\right)C$$

$$= 32.5 + \left(\frac{-4}{60}\right)5$$

$$= 32.5 - .33$$

$$= 32.2 \text{ years}$$

Standard Deviation:

$$\sigma = C\sqrt{\frac{\Sigma f(d'^2)}{n} - \left(\frac{\Sigma fd'}{n}\right)^2}$$

$$= 5\sqrt{\frac{104}{60} - \left(\frac{-4}{60}\right)^2}$$

$$= 6.6 \text{ years}$$

Let us examine the standard deviation formula used in Table 3-4. The first term under the square root, $\sqrt{\Sigma f(d'^2)/n}$, compares with the $\sqrt{\Sigma(x^2)/n}$ formula when deviations are computed from the actual mean. But there is an additional term under the square root $(\Sigma fd'/n)^2$, similar in part to the term added in the computation of the

---

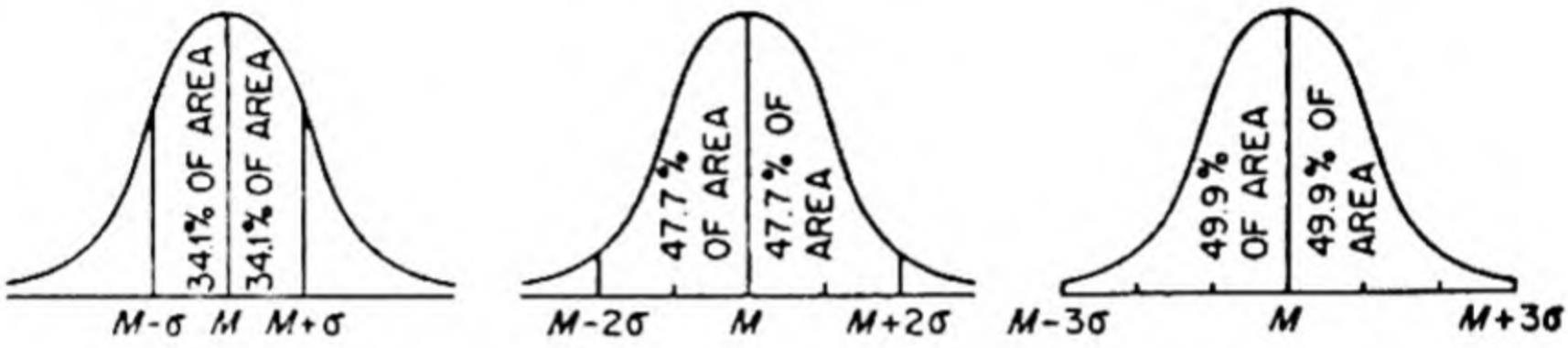
For Grouped Data:

$$\sigma = \sqrt{\frac{\Sigma f(X^2)}{n} - \left(\frac{\Sigma fX}{n}\right)^2} \quad \text{or} \quad \sigma = \frac{1}{n}\sqrt{n\Sigma f(X^2) - (\Sigma fX)^2}$$

Note the large X's. These formulas are convenient for computations with a calculator.

mean from the guessed mean, to correct in each case for the fact that deviations are from a guessed and not necessarily actual mean. Note that only one new column is needed to compute the standard deviation beyond that needed for the mean.

The mean age of the family clan is 32.2 years and the standard deviation, 6.6 years. The standard deviation is important in statistical theory. We shall use it in later chapters to measure area under the normal curve. About two-thirds of the area under a normal curve will lie within one standard deviation on each side of the mean, about 95 per cent will lie within two standard deviations on either side of the mean, and more than 99 per cent, within three standard deviations from the mean.



Area under a normal curve within one standard deviation from the mean.

Area under a normal curve within two standard deviations from the mean.

Area under a normal curve within three standard deviations from the mean.

Fig. 3.2.

One measure of dispersion *around the median* is the semi-interquartile range (Q). It is the mean distance of the first and third quartiles from the median. If Q_1, the first quartile, is the value below which 25 per cent of the total frequency lies, Q_2, the second quartile or median, is the value below which 50 per cent of the cases lie and Q_3, the third quartile, is the value below which 75 per cent of the total cases lie, then the semi-interquartile range Q is equal to:

$$Q = \frac{(Q_3 - Q_2) + (Q_2 - Q_1)}{2} = \frac{(Q_3 - Q_1)}{2} \qquad (10)$$

The semi-interquartile range takes into account only the middle half of the data, the data between Q_3 and Q_1. If Q is great, the dispersion of the inside cluster must be great. With a symmetrical distribution. $Q_3 - Q_2 = Q_2 - Q_1$. That is, 25 per cent of the frequency above and below the median is equidistant from the median.

If, in a test in mathematics, 75 per cent of the students received grades below 85, the median grade was 80, and 25 per cent were below 70, then the semi-interquartile range would be 7.5:

$$Q = \frac{(85 - 80) + (80 - 70)}{2} = \frac{15}{2} = 7.5$$

which is the mean distance of the first and third quartiles from the median. This distribution of grades is not symmetrical. It has the same per cent of cases 5 points above the median that it has 10 points below the median.

The first quartile Q_1 is computed in the same manner as the median except that we count up to the interval containing $\frac{1}{4}\,n$ and not to $\frac{1}{2}\,n$. To determine the value of Q_3, the third quartile, we count up to the interval containing $\frac{3}{4}\,n$.[2]

KEY TERMS

mean deviation **semi-interquartile range** **standard deviation**
range **variance**

REFERENCES

Dixon, W. J., and F. J. Massey, *Introduction to Statistical Analysis*, chap. 3. New York: McGraw-Hill Book Company, 1951.

Freund, J. E., *Modern Elementary Statistics*, chap. 5. New York: Prentice-Hall, Inc., 1952.

Lindquist, E. F., *A First Course in Statistics*, chap. 6. Boston: Houghton Mifflin Company, 1942.

Mode, Elmer B., *Elements of Statistics*, rev. ed., chap. 3. New York: Prentice-Hall, Inc., 1951.

Peatman, J. G., *Descriptive and Sampling Statistics*, chap. 7. New York: Harper and Bros., 1947.

EXERCISES

1. Compute the mean deviation and the standard deviation for the two income groups of five and seven members given below:

I. \$4,000, \$4,200, \$4,400, \$4,600, \$4,800
II. \$3,000, \$4,000, \$4,200, \$4,400, \$4,600, \$4,800, \$5,800

Which measure of dispersion, the mean deviation or the standard deviation, appears to be more affected by extreme deviations?

[2] The semi-interquartile range is rarely used as a measure of dispersion. The most common measures are the standard deviation, the variance, and the range.

2. In a scale for measuring attitudes toward war, there are 20 items, each ranked according to its expression of pro- or anti-war sentiment. A pro-war statement would be ranked close to 1 or 2 and an anti-war statement, close to 19 or 20.

Each individual taking the attitude test is asked to check the items with which he agrees. His attitude score is the median value of the items he checks.

In a class of 100 students, the scores for individuals extend from 1 to 18. The distribution is as follows:

Attitude Scores	*Frequency*
1– 2	5
3– 4	10
5– 6	11
7– 8	17
9–10	16
11–12	15
13–14	11
15–16	9
17–18	6

(a) Compute the mean and standard deviation for this distribution of attitude scores.

(b) Draw a histogram and mark on it the mean and plus and minus 1 standard deviation.

3. If the first quartile of age of female at first marriage is 22 years, the second is 24, and the third, 28, what is the semi-interquartile range? Is the distribution symmetrical? Why or why not?

4. Does the small sample below indicate any relationship between intelligence and ethnocentrism? (Intelligence is measured by the Wechsler-Bellevue Intelligence test; ethnocentrism, by ten items on an ethnocentrism scale.) Is there an apparent error on the table?

Mean Wechsler-Bellevue IQ Scores for Each Quartile on the Ethnocentrism Scale
(Psychiatric clinic, men and women)

Ethnocentrism Scale Quartiles	*Range*	*Number*	*Mean IQ*
Low quartile	10–24	8	125.3
Low-middle quartile	25–36	5	117.8
High-middle quartile	37–50	13	113.9
High quartile	51–70	11	107.3

SOURCE: Daniel Levinson, "Ethnocentrism in Relation to Intelligence and Education," in *The Authoritarian Personality*, Adorno, Frenkel-Brunswik, Levinson, and Sanford (New York: Harper and Bros., 1950), p. 283.

CHAPTER 4

USE OF THEORETICAL DISTRIBUTIONS AS MATHEMATICAL MODELS

4.1. The Normal Curve

Problem. We want to compare the variability in radicalism-conservatism between freshmen college students in an urban industrial area and students from a nearby rural college.

Method. A radicalism-conservatism attitude test is given to an approximately representative sample of 100 freshmen students in each of the two colleges. There are thirty items in the test. The most radical statements are ranked close to 1 or 2, the most conservative statements, close to 29 or 30. Each student taking the attitude test is asked to check the items with which he agrees. His attitude score is the median value of the items he checks.

Table 4-1 gives the frequency distributions of student attitude scores among the two groups of freshmen students, and Figs. 4.1a and 4.1b give the histograms for the attitude scores. These histograms are the observed distributions. Their irregularities can be ironed out until they become smooth curves, similar to the kind we would obtain if the number of items on the test and the number of students taking the test were increased indefinitely.

Histograms, even when smoothed, may be of various shapes. For example, they may be:

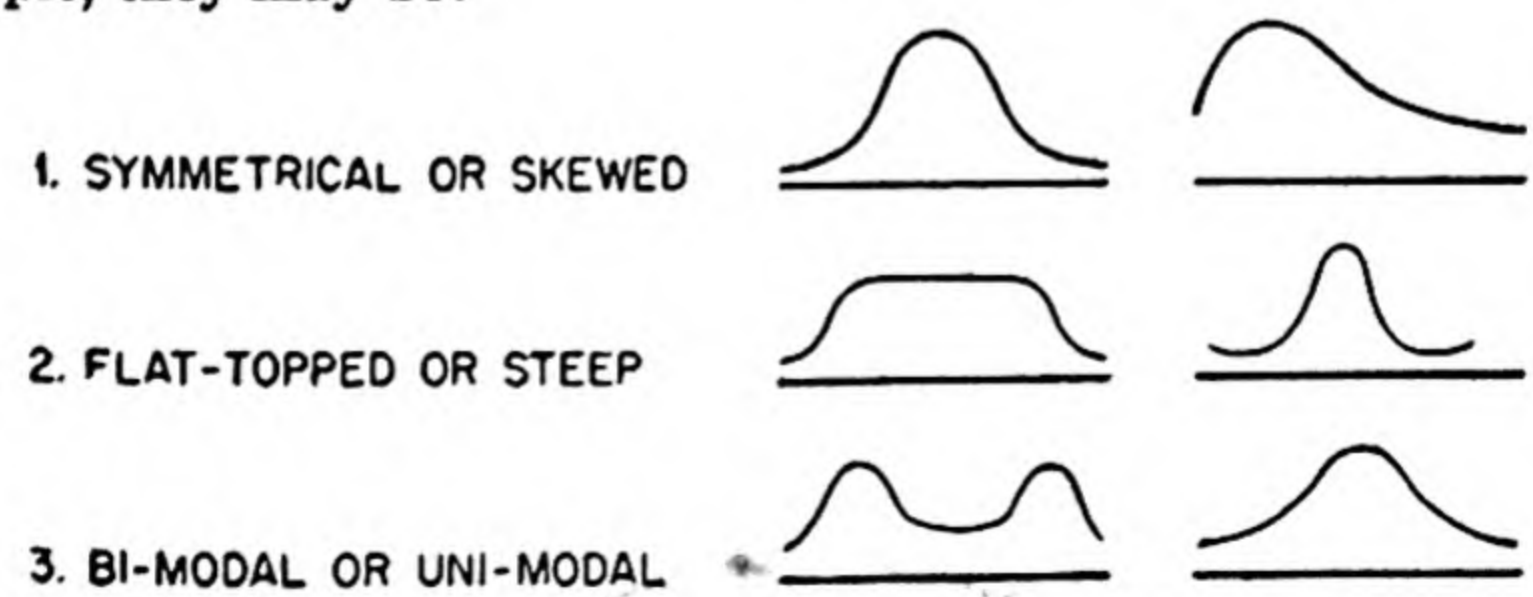


Table 4-1. Computation of Mean and Standard Deviation for Scores in Radicalism-Conservatism Test Given to Sample of Freshmen at Two Colleges
(Hypothetical data)

	Freshmen in Urban College					Freshmen in Rural College				
Scores	*Interval Limits*	*Frequency*	d'	fd'	$f(d'^2)$	*Scores*	*Interval Limits*	*Frequency*	d'	fd'	$f(d'^2)$
1- 2	.5- 2.4*	4	−7	−28	196	1- 2	.5- 2.4*	0	−7	0	0
3- 4	2.5- 4.4	4	−6	−24	144	3- 4	2.5- 4.4	1	−6	−6	36
5- 6	4.5- 6.4	6	−5	−30	150	5- 6	4.5- 6.4	2	−5	−10	50
7- 8	6.5- 8.4	9	−4	−36	144	7- 8	6.5- 8.4	4	−4	−16	64
9–10	8.5–10.4	9	−3	−27	81	9–10	8.5–10.4	5	−3	−15	45
11–12	10.5–12.4	11	−2	−22	44	11–12	10.5–12.4	7	−2	−14	28
13–14	12.5–14.4	13	−1	−13	13	13–14	12.5–14.4	8	−1	−8	8
15–16	14.5–16.4	14	0	0	0	15–16	14.5–16.4	10	0	0	0
17–18	16.5–18.4	9	1	9	9	17–18	16.5–18.4	12	1	12	12
19–20	18.5–20.4	7	2	14	28	19–20	18.5–20.4	14	2	28	56
21–22	20.5–22.4	5	3	15	45	21–22	20.5–22.4	11	3	33	99
23–24	22.5–24.4	4	4	16	64	23–24	22.5–24.4	10	4	40	160
25–26	24.5–26.4	3	5	15	75	25–26	24.5–26.4	7	5	35	175
27–28	26.5–28.4	1	6	6	36	27–28	26.5–28.4	5	6	30	180
29–30	28.5–30.4	1	7	7	49	29–30	28.5–30.4	4	7	28	196
Sum		100 (= n)		−98	1,078	Sum		100 (= n)		137	1,109

Mean = Guessed mean + $\left(\frac{\Sigma fd'}{n}\right) C$

= 15.5 + $\left(\frac{-98}{100}\right) 2$

= 13.5

Standard Deviation = $C\sqrt{\frac{\Sigma f(d'^2)}{n} - \left(\frac{\Sigma fd'}{n}\right)^2}$

= $2\sqrt{\frac{1078}{100} - \left(\frac{-98}{100}\right)^2}$

= $2\sqrt{10.78 - .96} = 2\sqrt{9.82}$

= ±6.3

Mean = Guessed mean + $\left(\frac{\Sigma fd'}{n}\right) C$

= 15.5 + $\left(\frac{137}{100}\right) 2$

= 18.2

Standard Deviation = $C\sqrt{\frac{\Sigma f(d'^2)}{n} - \left(\frac{\Sigma fd'}{n}\right)^2}$

= $2\sqrt{\frac{1109}{100} - (1.37)^2}$

= $2\sqrt{11.09 - 1.88} = 2\sqrt{9.21}$

= ±6.1

* That is, up to, but not including 2.5.

The normal curve is one form of curve frequently used to smooth out sociological data. The normal curve can be used only if the observed data approximate a normal distribution. To discover

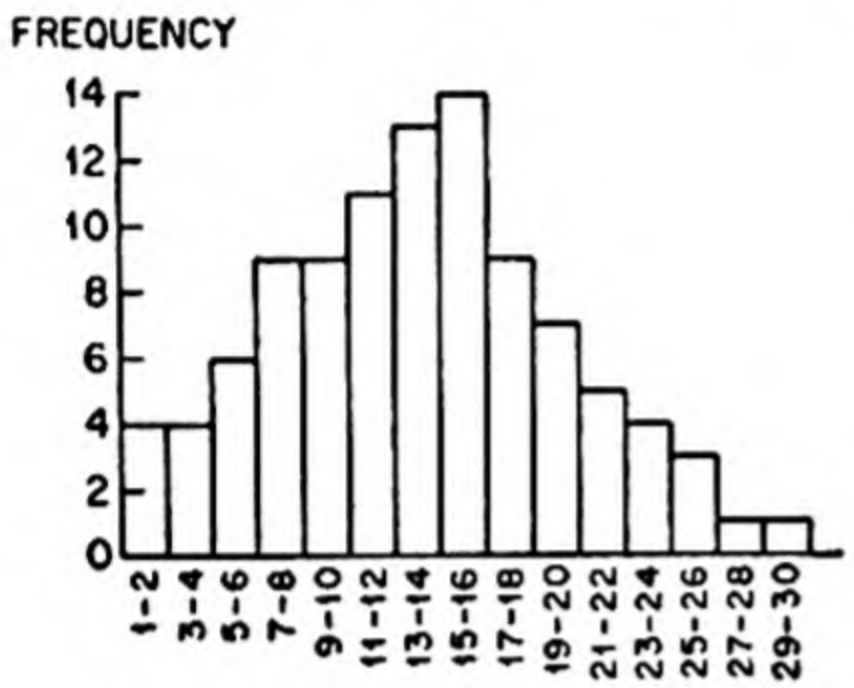


Fig. 4.1a. Scores in radicalism-conservatism test given to 100 students in hypothetical urban college.

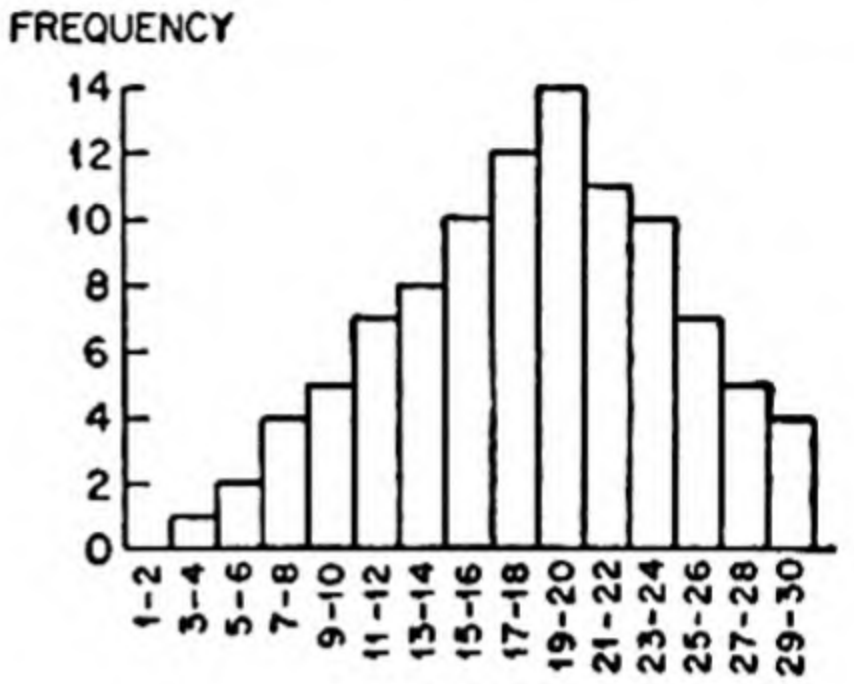
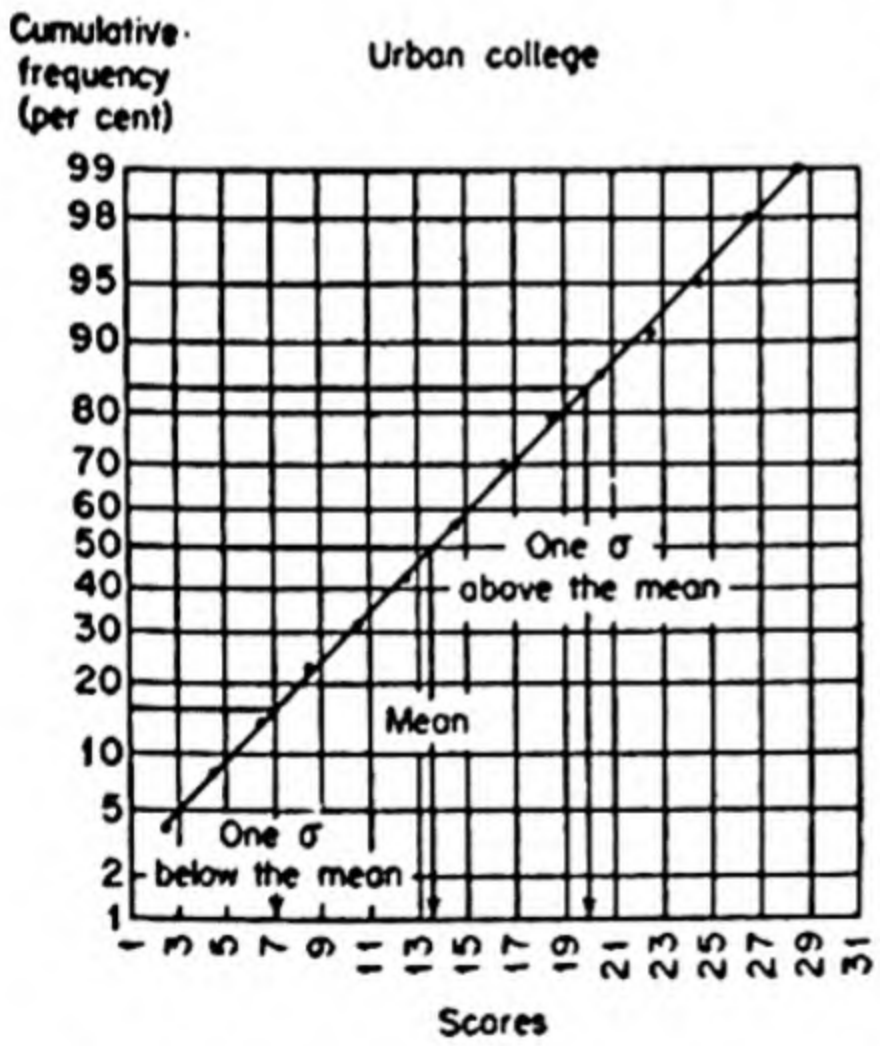


Fig. 4.1b. Scores in radicalism-conservatism test given to 100 students in hypothetical rural college.

whether our two observed distributions approximate normal distributions, we shall plot their cumulative frequencies on normal-curve graph paper (Fig. 4.2). On the horizontal scale of the paper are



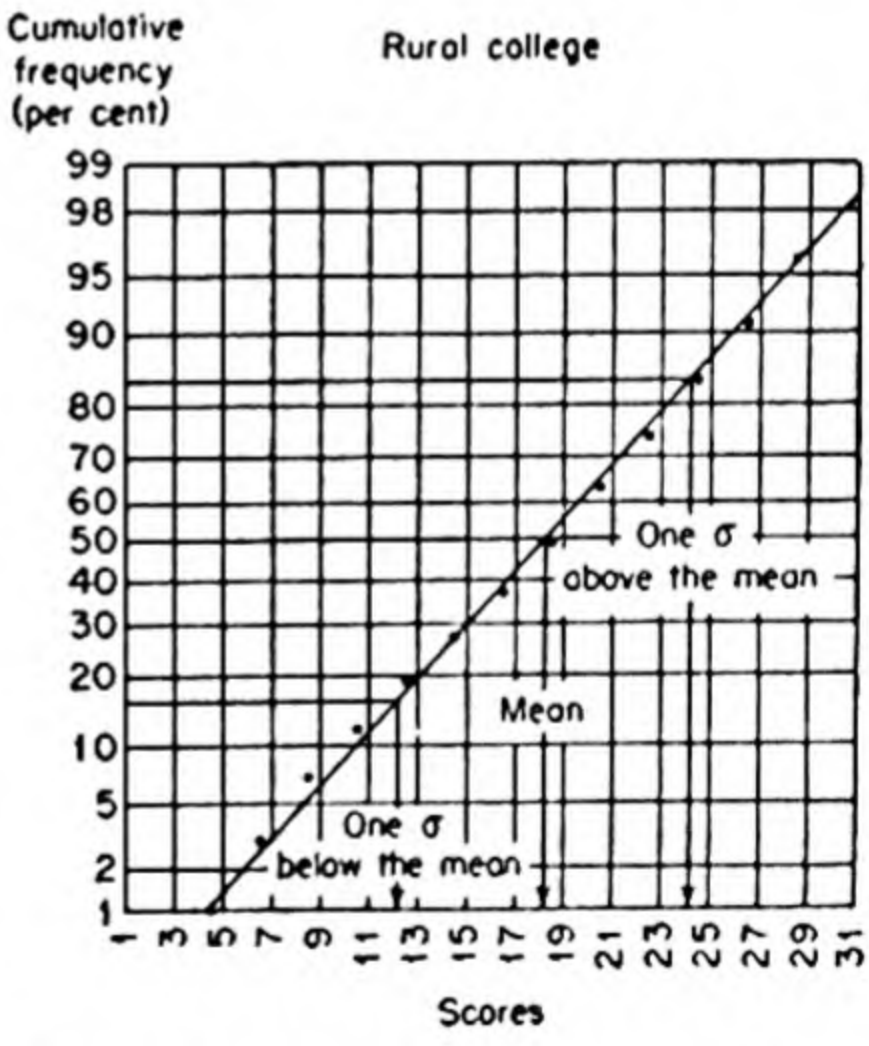


Fig. 4.2. Cumulative distribution of radicalism-conservatism scores at hypothetical urban and rural colleges plotted on normal-curve graph paper.

Table 4-2. Frequency and Cumulative Frequency Distribution of Scores in Hypothetical Urban and Rural College
(In percentages)

Urban College				Rural College			
Scores	*Interval Limits*	*Frequency*	*Cumulative Frequency*	*Scores*	*Interval Limits*	*Frequency*	*Cumulative Frequency*
1– 2	.5– 2.4	4	4	1– 2	.5– 2.4	0	0
3– 4	2.5– 4.4	4	8	3– 4	2.5– 4.4	1	1
5– 6	4.5– 6.4	6	14	5– 6	4.5– 6.4	2	3
7– 8	6.5– 8.4	9	23	7– 8	6.5– 8.4	4	7
9–10	8.5–10.4	9	32	9–10	8.5–10.4	5	12
11–12	10.5–12.4	11	43	11–12	10.5–12.4	7	19
13–14	12.5–14.4	13	56	13–14	12.5–14.4	8	27
15–16	14.5–16.4	14	70	15–16	14.5–16.4	10	37
17–18	16.5–18.4	9	79	17–18	16.5–18.4	12	49
19–20	18.5–20.4	7	86	19–20	18.5–20.4	14	63
21–22	20.5–22.4	5	91	21–22	20.5–22.4	11	74
23–24	22.5–24.4	4	95	23–24	22.5–24.4	10	84
25–26	24.5–26.4	3	98	25–26	24.5–26.4	7	91
27–28	26.5–28.4	1	99	27–28	26.5–28.4	5	96
29–30	28.5–30.4	1	100	29–30	28.5–30.4	4	100

placed the scores; on the vertical scale, the cumulative frequency percentages. The vertical scale of the graph paper is stretched in such a fashion that the cumulative frequency distribution will fall along a straight line if the scores are normally distributed.

According to Fig. 4.2, the distributions make approximately straight lines on the normal-curve graph paper. Assuming that our data are normal, we shall use the normal curve to represent our populations.[1]

Description of the Normal Curve. The concept of the normal curve was originally developed in a mathematical treatise by De Moivre in 1733, who considered its application in problems encountered in games of chance.

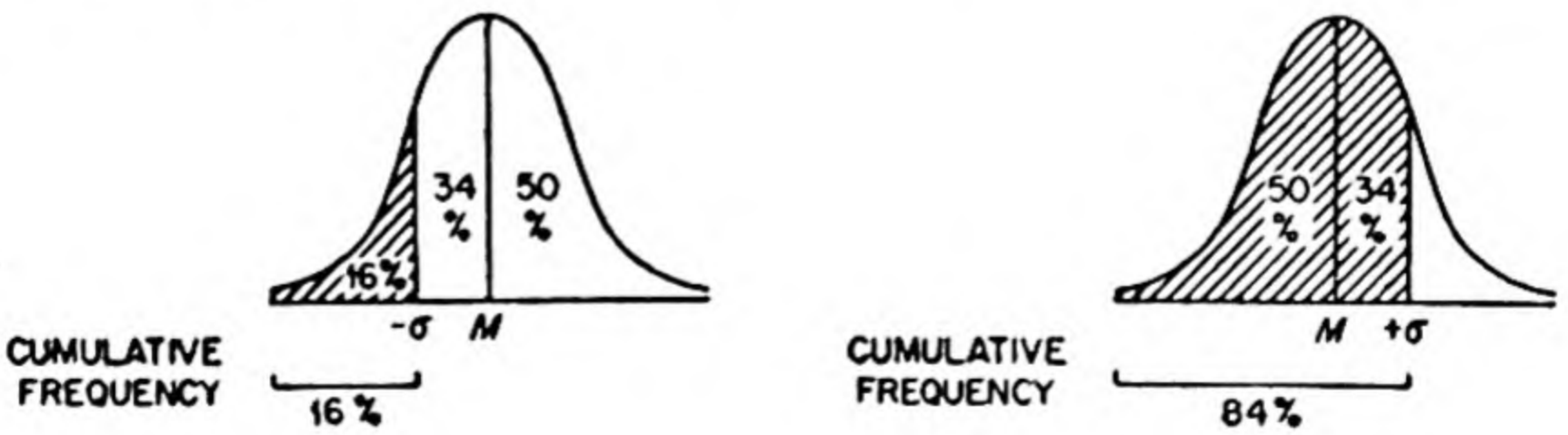
The term "normal" does not signify the customary or the most frequent type of a distribution of a variable, but refers to the form of the distribution. A few of the characteristics of the normal curve, and of the normal distribution are:

(1) The normal curve is symmetrical about the mean. The number of cases below the mean in a normal distribution are equal to the number of cases above the mean, which makes the mean and the median coincide. The height of the curve for a positive deviation of 10 units is the same as the height of the curve for a negative deviation of 10 units.

(2) The height of the normal curve is at its maximum at the mean. Hence the mean and the mode of the normal distribution coincide.

(3) There is one maximum point of the normal curve, which occurs

[1] If the distribution does approximate normality, we can read off from the graph the approximate mean and standard deviation. The mean score of a normal distribution has a cumulative frequency of 50 per cent; the score that is one standard deviation below the mean has a cumulative frequency of 16 per cent (50% minus 34%); and the score that is one standard deviation above the mean has a cumulative frequency of 84 per cent (50% plus 34%). Hence the mean score at the urban college is approximately 13.5, the score one standard deviation below the mean approximates 7, and the score one standard deviation above the mean is close to 20.



at the mean. The height of the curve declines as we go in either direction from the mean. This dropping off is slow at first, then rapid (as the curve starts bending outward), then slow again. Theoretically, the curve never touches the base line. Its tails approach, but never reach the horizontal line. Hence its range is unlimited.

(4) The variable distributed according to the normal curve is a continuous and not a discrete variable. The use of the normal curve to smooth out the distribution of the discrete variable, radicalism-conservatism scores, represents, at best, only an approximation. We assume that a score of 1 covers the interval .50 to 1.49.

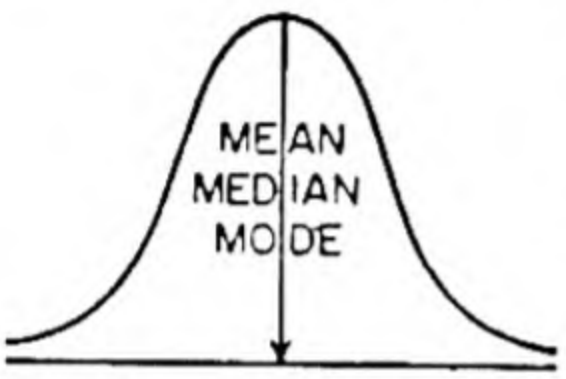


Fig. 4.3. The normal curve.

Application of the Normal Curve to Our Observed Distributions. We shall define a radical as one who has a score of 10 or below in the radicalism-conservatism test. What per cent of the freshmen students in the two schools are radical, assuming that our 100 freshmen are representative of total freshmen students? We could estimate these percentages from the sample histograms themselves. However, if the total freshman population is a normal one, the smooth, normal curve should better represent this population than one sample of 100 students.

The histograms of Figs. 4.1a and 4.1b are reproduced in Fig. 4.4. Upon the histograms are superimposed normal curves.

Fig. 4.4. Normal curves superimposed upon histograms of Figs. 4.1a and 4.1b.

Below the raw score scale on the histograms of Fig. 4.4, we have drawn a standard scale, that is, a scale in standard deviation units. The 0 point of the standard scale is placed at the mean of the raw score values, which, in our two distributions, are 13.5 and 18.2. The 0 at the center of the curve signifies that this point is 0 deviations away from the mean. It is at the highest point of the normal curve, just at the mean.

The plus 1 point on the standard scale is placed 1 standard deviation above the mean. For the urban school it is 13.5 plus 6.3, for the rural school, 18.2 plus 6.1. It is at this point, 1 standard deviation away from the mean, that the curve stops bending inward and starts bending outward. We continue in a similar fashion marking off minus 1 standard deviation, plus and minus 2 standard deviations, plus and minus 3 standard deviations (if our observed distributions extend out this far).

We shall not explain the method of fitting the theoretical normal curve to an observed distribution. (Several references at the end of the chapter indicate the procedure for fitting the normal curve.) Although a normal curve is implicitly being fitted to observed data whenever we use the normal curve area table, it is not necessary actually to superimpose the curve on the data.

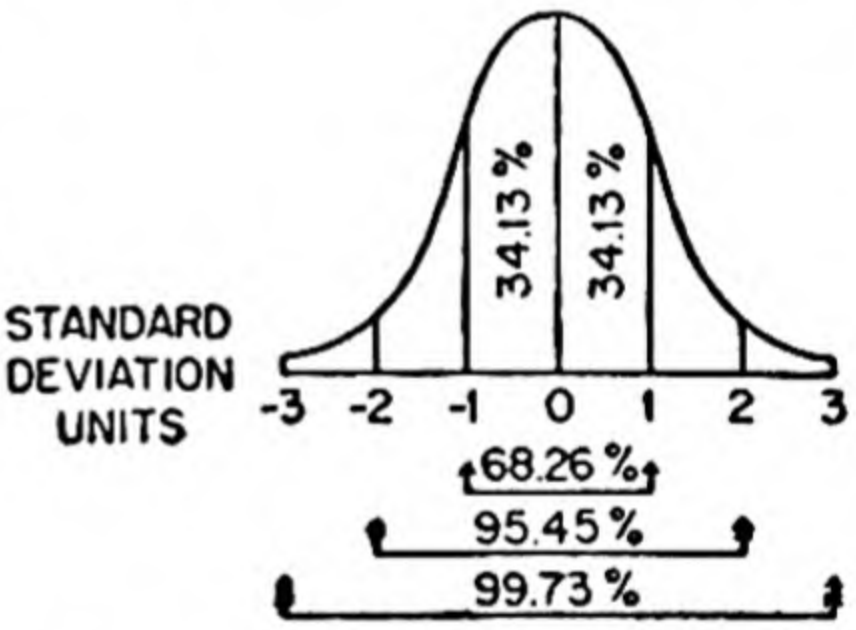


Fig. 4.5. Area under the normal curve within one, two, and three standard deviations of the mean.

Normal curves may differ among themselves in their *means*, which tell where to place the center of the curve, and in their *standard deviations*, which tell how widely to spread the curve. The area distribution, however, is always the same. The area included between the mean and the point on the base line one standard deviation *above the mean* is 34.13 per cent of the total, and the area between the mean and one standard deviation *below the mean* is 34.13 per cent; hence 68.26 per cent of the total area under the normal curve lies within one standard deviation of the mean. The two standard deviation limits (plus and minus) will include 95.45 per cent of the area under the curve; plus and minus three standard deviations, 99.73 per cent.

Theoretically, the range of the normal curve is unlimited, but most observations will usually fall within three standard deviations, and nearly all within four standard deviations.

For a *normally distributed* variable, the *relative frequency*, or probability, of observations between any two points on the horizontal scale is equal to the *area* under the normal curve between these two points. Note in Fig. 4.4 that the frequency of observations within one standard deviation in the observed distributions is not exactly 68.26 per cent, which is the area under the normal curve within one standard deviation of the mean. The observed distributions follow only approximately, and not precisely, the theoretical normal curve distribution.

Having fitted the normal curve to our observed distributions, we want to determine what per cent of a normal population at both colleges are radical. A radical is defined as a person who has made a score of 10 or below in the radicalism-conservatism test. To apply normal curve distributions, this raw score of 10 must be converted into a standard score in which distances are measured in standard deviation units. We have assumed that the discrete score 10 is continuous from 9.50 to 10.50. To determine what proportion of the area under the normal curve is at 10 or below, we shall compute the standard score for 10.5. The raw score of 10.5 can be converted into a standard score by calculating the deviation from the mean of its distribution and dividing by its standard deviation.

Standard Score:
$$\frac{X - \bar{X}}{s} = \frac{x}{s} \tag{11}$$

Converting a raw score of 10.5 into a standard score:

Urban College:
$$\frac{X - \bar{X}}{s} = \frac{10.5 - 13.5}{6.3} = \frac{-3.0}{6.3} = -.48 \text{ standard deviation units}$$

Rural College:
$$\frac{X - \bar{X}}{s} = \frac{10.5 - 18.2}{6.1} = \frac{-7.7}{6.1} = -1.3 \text{ standard deviation units}$$

At the urban college, a raw score of 10.5 is equivalent to a standard score of $-.48$. If radicalism-conservatism is a normally distributed variable at the urban college, we want to find out what per cent of the students are radical, i.e., what per cent have a raw score lower than

10.5, or an equivalent standard score lower than −.48. To find out what per cent of the students are radical, we must first determine where −.48 standard deviation units lie under the normal curve.

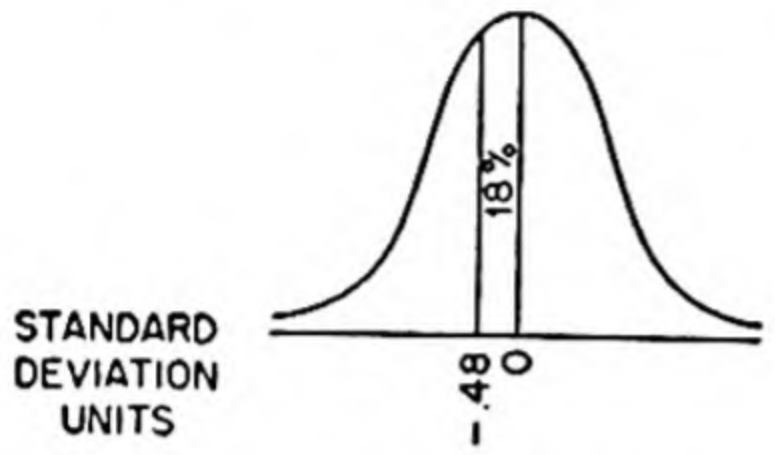


Fig. 4.6a. Per cent of area under normal curve between mean and −.48 standard deviation units below the mean.

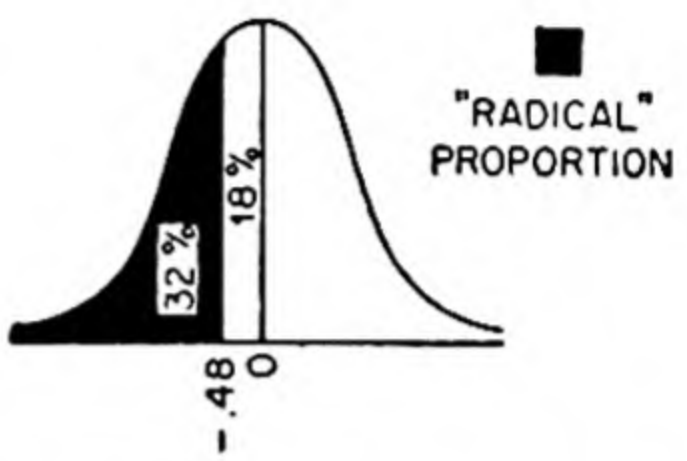


Fig. 4.6b. Per cent of area under normal curve below − 48 standard deviation units.

This information is given in Table II of the Appendix. The *left-hand column* of Table II gives *distances from the mean of the normal curve in standard deviation units* (x/σ or x/s), the top row adding the second decimal place to the standard-deviation distance from the mean. The *body* of Table II gives the *area under the normal curve* for each specified standard-deviation distance from the mean. The area is given for positive values of the standard scores. Because of the symmetry of the normal curve, the area under the normal curve between the mean and a specified standard score is the same whether the score is positive or negative.

Table II indicates that the distance from the mean to 0.48 standard deviations below the mean covers 18 per cent of the total area of the curve. But we want to know not the area between the mean and −.48 standard deviations, but the area *below* −.48 standard deviations. We find this out only indirectly from Table II. Fifty per cent of the area of the curve is below the mean. Hence the area of the curve *below* a standard score of −.48 is 50 per cent minus 18 per cent, or 32 per cent. If radicalism-conservatism is a normally distributed variable among the students of the urban college, 32 per cent of the freshmen students are radical.

At the rural college, a raw score of 10.5 is equivalent to a standard score of −1.3. The area under the curve from the mean to 1.3 standard deviations below the mean is 40 per cent. Consequently, the area under the curve below −1.3 standard deviations is 50 per

cent minus 40 per cent, or 10 per cent. If radicalism-conservatism is a normally distributed variable among the students of the rural college, 10 per cent of the students are radical.

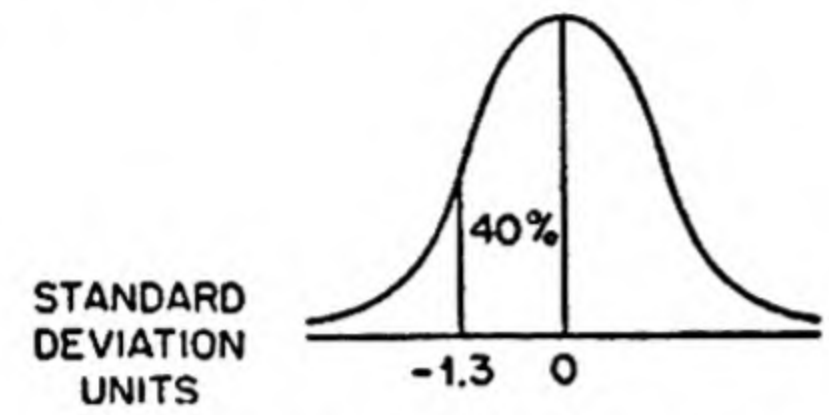


Fig. 4.7a. Per cent of area under normal curve between the mean and 1.3 standard deviation units below the mean.

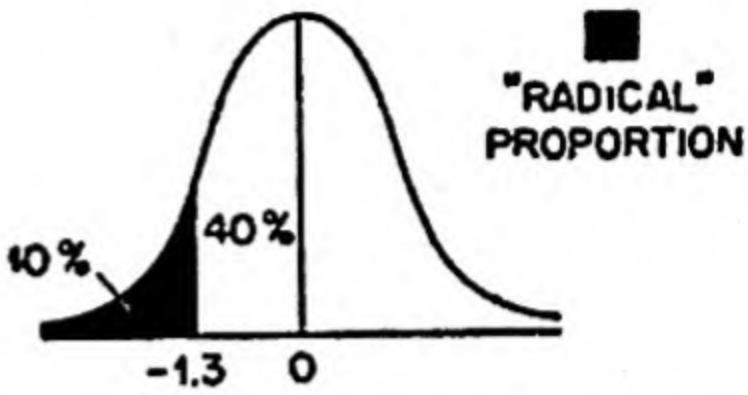


Fig. 4.7b. Per cent of area under normal curve below 1.3 standard deviation units.

Is Sociological Data Normally Distributed? Many physical traits such as height and weight may show an approximately normal distribution for a limited range if the population is highly homogeneous with regard to certain related traits, for example, the same sex and the same race. But even such physical measures as height and weight cannot precisely follow the normal curve. The range of height and weight is not unlimited. And it is doubtful whether the mean and median weight would also be the modal weight.

Tests are often deliberately constructed, through such means as adjustment of the difficulty of the question, to give an approximately normal distribution of scores. Grades, too, may be assigned to conform to a normal distribution. For example, the interval $\bar{X} \pm 2.5$ standard deviations can be divided into five equal parts corresponding to the grades A, B, C, D, and E, the proportion of students given each letter grade being determined by the area under the normal curve within each interval. Seven per cent of the area under the normal

Table 4-3. Grades Assigned to Conform to Normal Curve Distribution

Grades	*Range of Interval under Normal Curve*			*Area under Normal Curve within Interval*
E	$\bar{X} - 2.5\sigma$	to	$\bar{X} - 1.5\sigma$	7%
D	$\bar{X} - 1.5\sigma$	to	$\bar{X} - 0.5\sigma$	24%
C	$\bar{X} - 0.5\sigma$	to	$\bar{X} + 0.5\sigma$	38%
B	$\bar{X} + 0.5\sigma$	to	$\bar{X} + 1.5\sigma$	24%
A	$\bar{X} + 1.5\sigma$	to	$\bar{X} + 2.5\sigma$	7%

curve is between +1.5 and +2.5 standard deviations. Consequently, the highest 7 per cent in the class will receive an "A". Twenty-four per cent of the area is between +.5 and +1.5; therefore 24 per cent will receive "B".

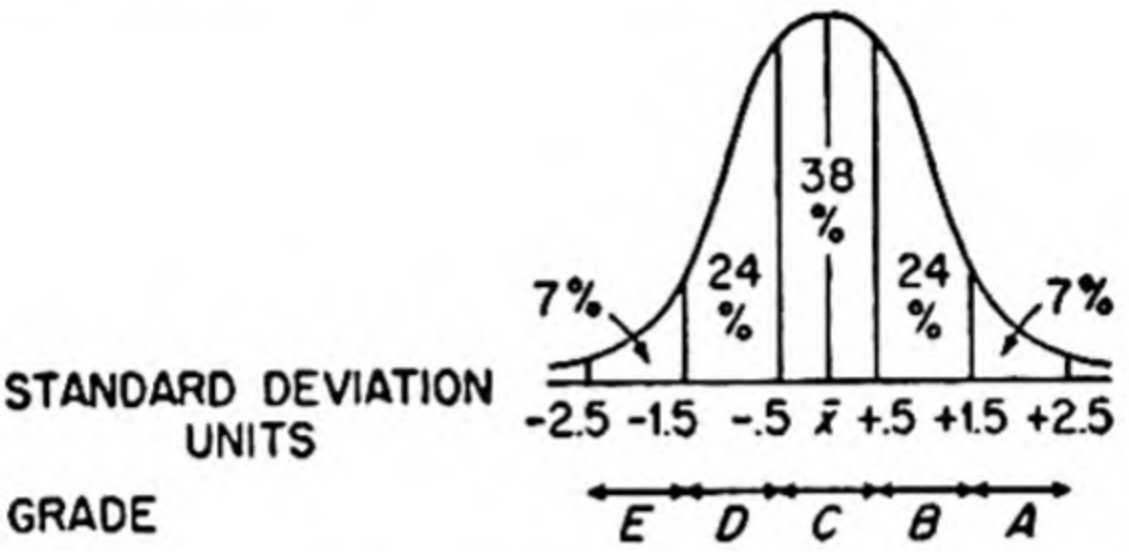


Fig. 4.8. Normal curve illustration for Table 4-3.

Variables that are not continuous cannot actually be normally distributed, since the intervals of the variable cannot conceivably get indefinitely small. This applies to many sociological variables: number of children per family, number of rooms per dwelling unit, intelligence quotients, attitude and opinion test scores, etc. It would also apply to the attitude test on radicalism-conservatism presented in this chapter. If it can be assumed, however, that the distribution of the variable tends to approach normality, the normal curve is often very useful as an approximation.

KEY TERMS

area under normal curve **standard score**

REFERENCES

Dixon, W. J., and F. J. Massey, *Introduction to Statistical Analysis*, chap. 5. New York: McGraw-Hill Book Company, 1951.

Freund, J. E., *Modern Elementary Statistics*, chap. 6. New York: Prentice-Hall, Inc., 1952.

Hagood, M. J., and D. O. Price, *Statistics for Sociologists*, chap. 14. New York: Henry Holt and Co., 1952.

Mode, Elmer, *Elements of Statistics*, Rev. ed., chap. 8. New York: Prentice-Hall, Inc., 1951.

Peatman, J. G., *Descriptive and Sampling Statistics*, chap. 8. New York: Harper and Bros., 1947.

EXERCISES

For each problem, draw a free-hand sketch of a normal curve and insert the information given.

1. If heights of college freshmen are found to be distributed normally, what is the probability that a randomly selected freshman ranks at, or above, plus 1 standard deviation in height? (*Note:* The statement that "The probability that a randomly selected freshman ranks at or above plus 1 standard deviation is equal to . . ." means the same as ". . . per cent of the area under the normal curve is at or above plus 1 standard deviation.")

2. On the Stanford-Binet intelligence test the mean score of those taking the test has been found to be very close to 100 and the standard deviation, very close to 16. If student *A* scores 148, what would his standard score be? (Use the formula x/σ to convert a raw score into a standard score.)

3. In a group of 20,000 college freshmen, scores on a test of general aptitude are found to be approximately normally distributed. The mean score is 150 and the standard deviation, 25. What is the approximate number of students whose score ranged from 100 to 175?
(*Steps:* (a) Convert the raw scores of 100 and 175 into standard scores.
(b) Determine the area under the normal curve between these two standard scores.
(c) Interpreting area as relative frequency, apply the resulting percentage to the total freshman group.)

4. In a normal distribution with a mean of 65 and a standard deviation of 15, about 95 per cent of the population would lie between which two points? (*Note:* If 95 per cent of total area lies on both sides of the symmetrical normal curve, what per cent lies on one side? The boundary of this area corresponds with what standard-deviation distance from the mean? What actual raw score distance from the mean?)

5. The scores on an attitude test taken by a sample of 150 wives of corporation executives were distributed as follows:

Scores	*Frequency*
50– 59	7
60– 69	8
70– 79	10
80– 89	18
90– 99	22
100–109	21
110–119	18
120–129	16
130–139	15
140–149	15

(a) Find the mean and the standard deviation.
(b) Draw a histogram for this distribution and mark on it the mean and plus and minus one and two standard deviations.

(c) What percentage of the scores should lie within 1, 2, and 3 standard deviations in a normal distribution? How does this compare with the actual sample percentages?

(d) Does the fact that the sample is not normal mean that it was necessarily obtained from a skewed, non-normal population?

(e) If the universe distribution is normal, what are the chances that a single score selected at random will lie above 80? (*Note:* What per cent of total scores are above 80?)

(f) In a normal distribution, 5 per cent of the scores will lie more than what standard-deviation distance and what raw score distance from the mean? (*Note:* If we are interested in a *total of 5 per cent* of the scores lying at the two ends of the curve, what per cent of these scores will lie at each end? What per cent of the area of the normal curve is between the mean and these scores? Look up this area in the body of the normal curve table to find out the corresponding standard score.)

4.2. The Binomial Distribution

Problem. Interviewers are sent out by a certain local agency to interview five families a day.

We want to determine the probability that, in a random selection of families, the interviewers will find no families at home, exactly one, two, three, four, and five families at home.

Method. One method of solving this problem is to have perhaps fifty interviewers visit five families and tabulate the relative frequency of families at home. Table 4-4 and Fig. 4.9 on page 64 provide the results of this observational evidence.

Assume, however, that either we cannot send these fifty interviewers out, or that we do not want to base our conclusions upon this sample.

If the call is made between three and five o·clock in the afternoon, it has been the agency's experience over a long period of years that in three calls out of five a family will be at home. If calls are made an indefinitely great number of times in the future under exactly similar conditions, we assume that the *probability* of a family being at home will remain three out of five or $\frac{3}{5}$.

We define *probability* as the relative frequency with which an event is expected to occur on the average, or in the long run.

The variable, family at home, is discrete and dichotomous. There are two possible events, and these two events are exhaustive and mutually exclusive. A family is either at home or not at home; it cannot both be at home and not be at home.

Let us call family at home a successful event; family not at home, an unsuccessful event.

Let p be the probability of a successful event at any single try (in this problem, at any ringing of a home door-bell), and let q be the probability of a failure at any single try.

Table 4-4. Frequency Distribution of Families at Home as Found by Visits of Fifty Interviewers to Five Families Each

Number of Successes (number of families at home)	Observed Frequency	
	Number	*Per cent*
0	1	2
1	5	10
2	10	20
3	16	32
4	13	26
5	5	10
	$n = 50$	100%

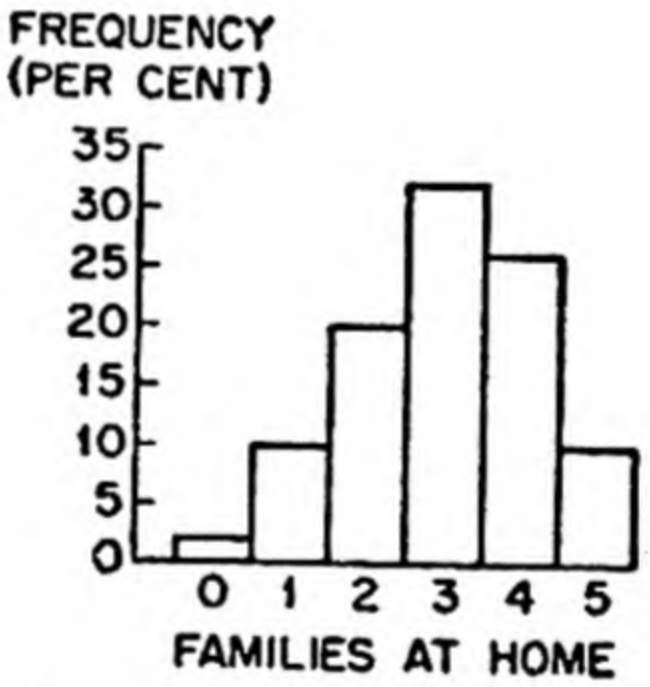


Fig. 4.9. Histogram for data of Table 4-4.

We want to know the probability of obtaining 0, 1, 2, 3, 4, and 5 successes out of 5 independent tries. By *independence* we mean that what happens at one door in no way influences or is related to what happens at any other door. The probability of one family being at home is the same whether or not another family is at home.

A family is considered to be at home if any resident over 16 years of age responds.

We have said that the terms, "family at home" and "family not at home," are mutually exclusive events. Two events are *mutually exclusive* if both cannot happen at the same time. The probability

of either of two mutually exclusive events occurring equals the sum of their probabilities:

$$p \quad \text{or} \quad q = p + q$$

If these two mutually exclusive events are *exhaustive*, the sum of their probabilities will equal one:

$$p + q = 1$$

A probability always lies between 0 and 1. There are no negative probabilities.

Our symbols are:

p = probability of success at any single try = $\frac{3}{5}$

q = probability of failure at any single try = $\frac{2}{5}$

p or $q = p + q = 1$

X = number of successes (0 or 1 or 2 or 3 or 4 or 5)

n = size of sample = 5 (number of tries)

Table 4-5, on page 66, shows how to determine the probability of 0, 1, 2, 3, 4, and 5 successes out of 5 tries.

If 5 families are visited, to determine the probability of exactly 2 families being at home (in Table 4-5, the row beginning with $X = 2$):

(1) We determine the probability of getting a combination of 2 families at home out of 5 independent family visits. If the visits are independent, i.e., the success or failure at one house in no way influences what happens at other houses, the probability of the occurrence of 2 successes and 3 failures equals the *product* of the probability that each will occur. Hence if the probability of success at any single try equals three-fifths and the probability of failure, two-fifths, then the probability of getting *a combination* of 2 families at home out of 5 family visits equals

$$p \cdot p \cdot q \cdot q \cdot q = p^2q^3 = \left(\frac{3}{5}\right)\left(\frac{3}{5}\right)\left(\frac{2}{5}\right)\left(\frac{2}{5}\right)\left(\frac{2}{5}\right) = \left(\frac{3}{5}\right)^2\left(\frac{2}{5}\right)^3$$

$$= \left(\frac{9}{25}\right)\left(\frac{8}{125}\right) = \frac{72}{3125}$$

(2) Next we compute the number of *different possible combinations* of 2 successes out of 5 tries. By *combinations* we mean here the num

Table 4-5. Probability Distribution of All Possible Number of Successes Out of Five Family Visits
$(p = \frac{3}{5})$

Number of Successes (number of families at home) (X) (1)	*Probability of Getting Any Single Combination* $p^X q^{n-X}$* Constant pb of Success $= \frac{3}{5}$ Constant pb of Failure $= \frac{2}{5}$ (2)	*Number of Different Possible Combinations of X Successes* $C_X^n = \frac{n!}{X!\,(n-X)!}$ (3)	*Probability of X Successes in 5 Independent Tries* $C_X^n p^X q^{n-X}$ (4) = (2) × (3)
0	0 successes, 5 failures $q\ q\ q\ q\ q = \left(\frac{2}{5}\right)^5 = \frac{32}{3125}$	$C_0^5 = \frac{5!}{0!\,5!} = 1$	$1 \times \frac{32}{3125} = \frac{32}{3125} = .01$
1	1 success, 4 failures $p\ q\ q\ q\ q = \left(\frac{3}{5}\right)\left(\frac{2}{5}\right)^4 = \frac{48}{3125}$	$C_1^5 = \frac{5!}{1!\,4!} = 5$	$5 \times \frac{48}{3125} = \frac{240}{3125} = .08$
2	2 successes, 3 failures $p\ p\ q\ q\ q = \left(\frac{3}{5}\right)^2\left(\frac{2}{5}\right)^3 = \frac{72}{3125}$	$C_2^5 = \frac{5!}{2!\,3!} = 10$	$10 \times \frac{72}{3125} = \frac{720}{3125} = .23$
3	3 successes, 2 failures $p\ p\ p\ q\ q = \left(\frac{3}{5}\right)^3\left(\frac{2}{5}\right)^2 = \frac{108}{3125}$	$C_3^5 = \frac{5!}{3!\,2!} = 10$	$10 \times \frac{108}{3125} = \frac{1080}{3125} = .34$
4	4 successes, 1 failure $p\ p\ p\ p\ q = \left(\frac{3}{5}\right)^4\left(\frac{2}{5}\right) = \frac{162}{3125}$	$C_4^5 = \frac{5!}{4!\,1!} = 5$	$5 \times \frac{162}{3125} = \frac{810}{3125} = .26$
5	5 successes, 0 failures $p\ p\ p\ p\ p = \left(\frac{3}{5}\right)^5 = \frac{243}{3125}$	$C_5^5 = \frac{5!}{5!\,0!} = 1$	$1 \times \frac{243}{3125} = \frac{243}{3125} = .08$

* Note that the exponent in each case is the number of successes and failures we are concerned with. *Total:* $\frac{3125}{3125} = 1.00$

ber of different ways of selecting 2 families out of 5 tries. For example, the first 2 families visited might have been at home, or the last 2, or the third and fourth. In all, there are 10 different possible combinations of 2 families at home out of 5 families visited.

Ways in Which Two Successes Can Be Selected Out of Five Family Visits

Families at Home

(1) Families 1 and 2
(2) Families 1 and 3
(3) Families 1 and 4
(4) Families 1 and 5
(5) Families 2 and 3
(6) Families 2 and 4
(7) Families 2 and 5
(8) Families 3 and 4
(9) Families 3 and 5
(10) Families 4 and 5

A formula for the determination of the number of different possible combinations of 2 successes out of 5 tries is:

$$C_X^n = \frac{n!}{X!\,(n-X)!} \tag{12}$$

$$= \frac{5!}{2!\,3!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 3 \cdot 2 \cdot 1}$$

$$= \frac{20}{2}$$

$$= 10$$

Symbols: C_X^n = number of different possible combinations or ways of selecting X successes out of n tries

n = size of sample (here, 5)

X = number of successes (here, 2)

$(n - X)$ = number of failures (here, 3)

$n!$, read n factorial, equals the product of all integers from n to 1, i.e., $(n)(n-1)(n-2)\cdots(2)(1)$

(0! is set equal to 1)

This combination formula can be used for any size sample, but becomes unwieldy where the sample is large. For example, try to work out the number of different possible combinations of 30 successes out of 50 tries.

(3) Finally, we multiply column (2) and column (3) of Table 4-4 to get the probability of 2 successful events in 5 independent tries ($\frac{720}{3125}$), or about 23 per cent. The chances are 23 out of 100 that exactly 2 families out of 5 will be at home. A formula for the probability of 2 successes out of 5 independent tries is:

$$C_X^n p^X q^{n-X} = 10\left(\frac{3}{5}\right)^2\left(\frac{2}{5}\right)^3 = 10\left(\frac{9}{25}\right)\left(\frac{8}{125}\right) = 10\left(\frac{72}{3125}\right)$$

$$= \frac{720}{3125} = .23 \qquad (13)$$

$$(X = 2, n = 5, \text{ and } p = \tfrac{3}{5})$$

Note that in Table 4-5 the sum of the 6 alternative probabilities is 1, or 100%. Either no families, or exactly 1, or 2, or 3, or 4, or 5 families are at home. The alternatives are mutually exclusive and exhaustive.

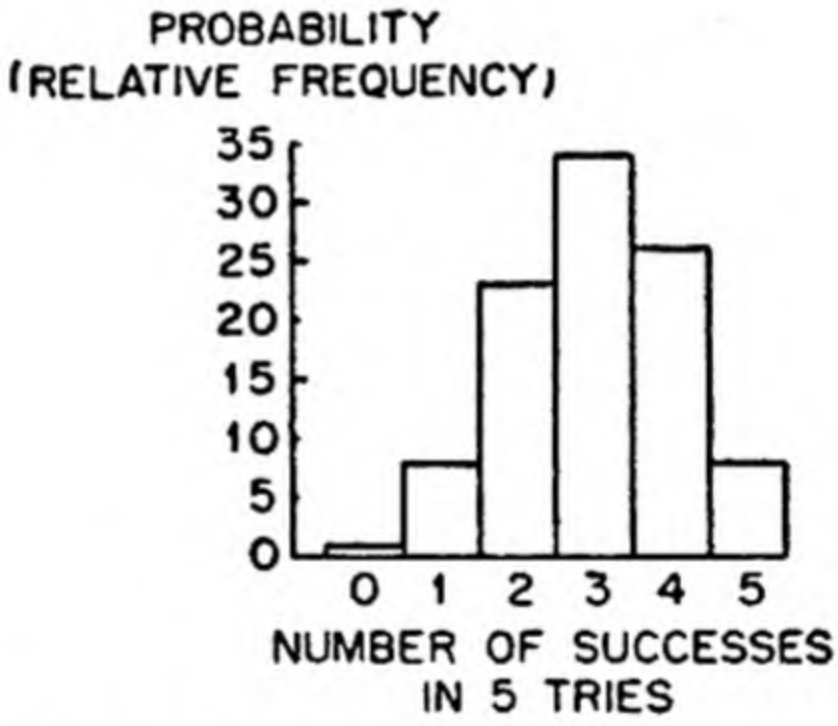


Fig. 4.10. Histogram for probability distribution of Table 4-5.

Figure 4.10 gives the histogram for the probability distribution of Table 4-5.

Theoretically then, if the constant probability of a family at home is $\frac{3}{5}$, there are 8 chances out of 100 that, in five family visits, exactly 5 families will be at home; the probability is 26 per cent that exactly 4 families will be at home, 34 per cent, that exactly 3 families will be at home. There is a probability of 91 per cent that *two or more* families will be at home. (See column 4, Table 4-5.)

Use of the Binomial in Computing the Probability of Getting Heads in the Toss of Coins The logic of the family-at-home example can be compared with that of the toss of a coin, where the probability of getting a head equals $\frac{1}{2}$, the probability of getting a tail equals $\frac{1}{2}$; we want to know the probability of getting 0, 1, 2, 3, 4, and 5 heads in the toss of 5 coins.

If we toss only one coin, the probability of getting a head equals $\frac{1}{2}$, the probability of getting a tail equals $\frac{1}{2}$.

If we *toss two coins*, the probability of getting 2 heads equals $\frac{1}{4}$ (combination 1), 2 tails equals $\frac{1}{4}$ (combination 4), one head and one tail equals $\frac{1}{2}$ (combinations 2 and 3). There are 4 combinations, and each combination is equiprobable (since p equals q equals $\frac{1}{2}$, p being the probability of getting a head and q, the probability of getting a tail.)

Four Different Combinations, Each Combination Equiprobable
$(p = q = \frac{1}{2})$

Combination	*First Coin*	*Second Coin*	*First Coin*	*followed by*	*Second Coin*
(1)	*H*	*H*			
(2)	*H*	*T*	*H*		*H*
(3)	*T*	*H*			*T*
(4)	*T*	*T*	or		
			T		*H*
					T

Table 4-6. Probability Distribution of All Possible Number of Heads in the Toss of Two Coins

(1)	(2)	(3)	(4)
Number of Heads (X)	*Probability of Getting Any Single Combination* $p^X q^{n-X}$	*Number of Different Possible Combinations*	*Probability of X Heads in Two Tries* (col. 2 × col. 3)
0	$(\frac{1}{2})^2 = \frac{1}{4}$	1	$\frac{1}{4}$
1	$(\frac{1}{2})(\frac{1}{2}) = \frac{1}{4}$	2	$\frac{1}{2}$
2	$(\frac{1}{2})^2 = \frac{1}{4}$	1	$\frac{1}{4}$
		Total: 4	$\frac{4}{4} = 1$

If we *toss three coins*, the probability of getting 3 heads equals $\frac{1}{8}$, 3 tails equals $\frac{1}{8}$, 2 heads and a tail equals $\frac{3}{8}$ (combinations 2, 3, and 5), 2 tails and a head equals $\frac{3}{8}$ (combinations 4, 6, and 7). There are 8 combinations, and each combination is equiprobable.

Eight Different Combinations, Each Combination Equiprobable
$(p = q = \frac{1}{2})$

Combination	*First Coin*	*Second Coin*	*Third Coin*
(1)	*H*	*H*	*H*
(2)	*H*	*H*	*T*
(3)	*H*	*T*	*H*
(4)	*H*	*T*	*T*
(5)	*T*	*H*	*H*
(6)	*T*	*H*	*T*
(7)	*T*	*T*	*H*
(8)	*T*	*T*	*T*

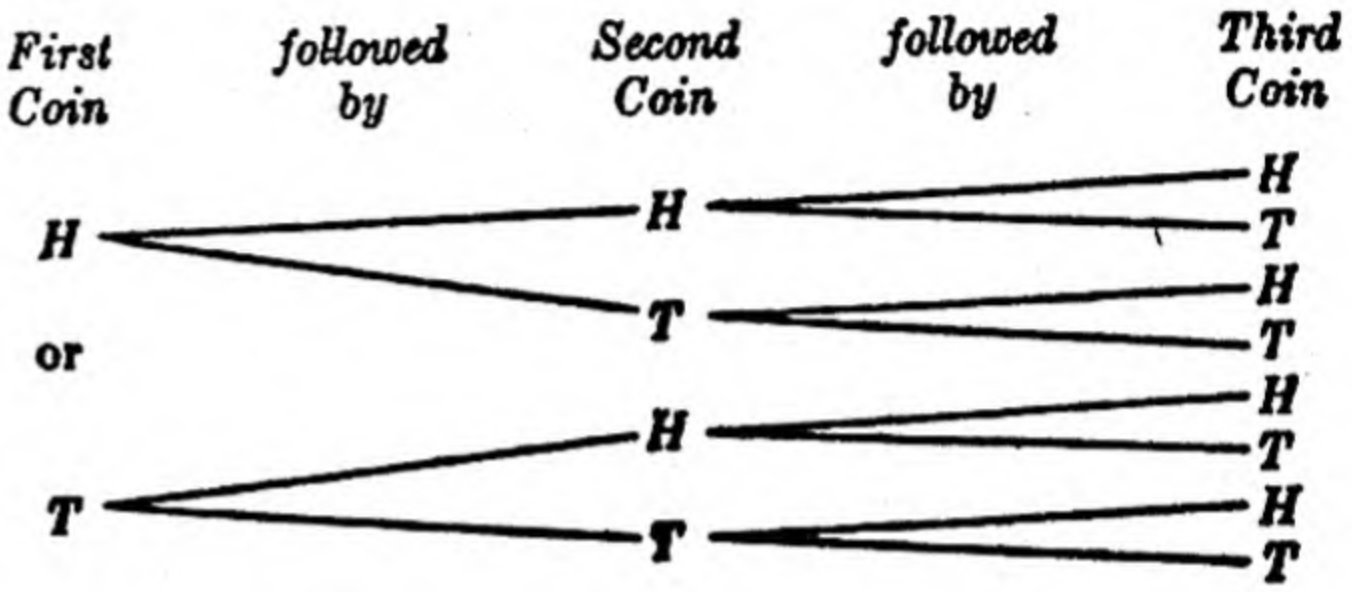


Table 4-7. Probability Distribution of All Possible Number of Heads in the Toss of Three Coins

(1) *Number of Heads* (X)	(2) *Probability of Getting Any Single Combination* $p^X q^{n-X}$	(3) *Number of Different Possible Combinations*	(4) *Probability of X Heads in Three Tries* (col. 2 × col. 3)
0	$(\frac{1}{2})^3 = \frac{1}{8}$	1	$\frac{1}{8}$
1	$(\frac{1}{2})(\frac{1}{2})^2 = \frac{1}{8}$	3	$\frac{3}{8}$
2	$(\frac{1}{2})^2(\frac{1}{2}) = \frac{1}{8}$	3	$\frac{3}{8}$
3	$(\frac{1}{2})^3 = \frac{1}{8}$	1	$\frac{1}{8}$
		Total: 8	$\frac{8}{8} = 1$

Each combination is equiprobable in the coin-tossing experiment because, with a perfect coin tossed on a perfect surface, the relative frequency of getting heads, in the long run, is $\frac{1}{2}$. Each combination is not equiprobable in the family-at-home problem because the probability of family-at-home is not equal to the probability of family-not-at-home.

The Normal Curve Approximation to the Binomial Distribution. The form of the binomial distribution depends on (1) the values of p and q and (2) the size of the sample (n). The distribution is symmetrical when $p = q = \frac{1}{2}$. In the four histograms of Fig. 4.11 the

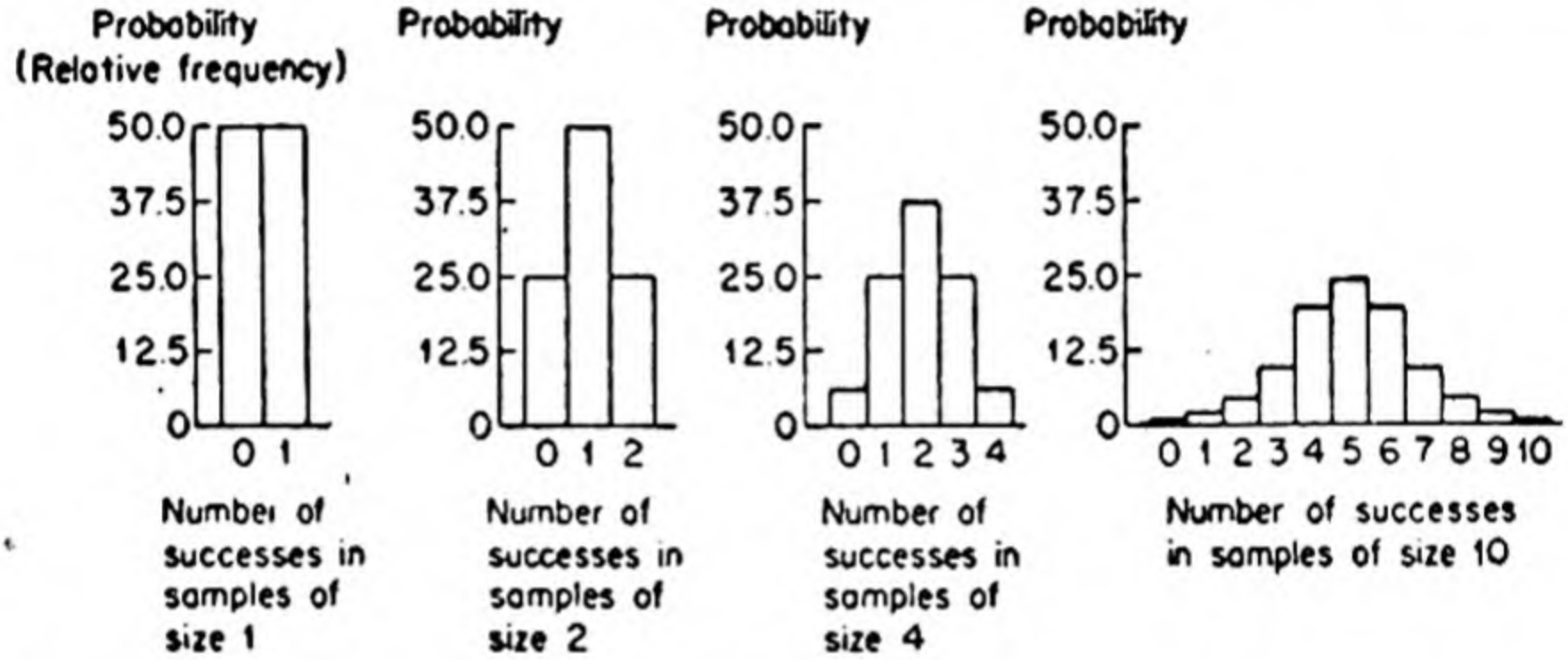


Fig. 4.11. Histograms for probability distributions of all possible numbers of successes in samples of size 1, 2, 4, and 10.

probability of success equals $\frac{1}{2}$ and the size of the sample is, consecutively, 1, 2, 4, and 10.

As the size of the sample increases, if p is not too close to 0 or 1, the binomial probability distribution looks more and more like the normal curve. Consequently, with large values of n, and p not too close to 0 or 1, the probabilities of the binomial distribution can be approximated by the normal distribution. This is a very useful approximation since, as the size of the sample gets increasingly large, the labor involved in the expansion of the binomial becomes prohibitive. The distribution of a discrete variable can never form a continuous curve, since it consists of a discrete set of points. But the sum of a number of binomial probabilities can be regarded as areas under the normal curve.

The problem of which normal curve to fit to a binomial distribution is determined by the mean and the standard deviation of the binomial

distribution. The formulas for the mean and standard deviation of the binomial distribution will be introduced in Chapter 7.

We have said that in a great number of repeated tries, a family will be at home in 3 calls out of 5. This probability equal to $\frac{3}{5}$ applies to a great number of repeated tries; it is the relative frequency in the long run. Can it be applied to estimate the probability that any single family chosen at random will be at home? In any single try a family will either be at home ($p = 1$) or will not be at home ($p = 0$). It cannot be $\frac{3}{5}$ at home. However, the probability equal to $\frac{3}{5}$ for a great number of repeated tries gives weight to the probability that any single family chosen at random is at home. We abbreviate this thinking by saying that the probability is $\frac{3}{5}$ that a "random" family will be at home.

When Can the Binomial Distribution Be Applied? The binomial distribution can be applied in sociological problems dealing with repeated trials of an event provided that:

1. The variable is discrete and dichotomous. For example, a family is either at home or not at home. The two categories are mutually exclusive.

2. All of the same relevant factors (and no new relevant factors) operating in the past will continue in the future. In our problem, the same economic and social conditions requiring a woman's presence in the home in mid-afternoon will continue to operate.

3. The trials are independent. What happens at one try has no influence upon, or relation to, succeeding tries.

4. We know from past experience the constant probability of success, that is, the probability of success at each separate try, or we can estimate this probability.

5. We have exhausted all possible events. The total probability of getting the possible events is equal to 1, or 100 per cent.

KEY TERMS

combinations **independent events** **probability**
exhaustive events **mutually exclusive events**

REFERENCES

Hagood, M. J., and D. O. Price, *Statistics for Sociologists*, chap. 15. New York: Henry Holt and Company, 1952.

Mode, Elmer, *Elements of Statistics*, Rev. ed., chap. 12. New York: Prentice-Hall, Inc., 1951.

EXERCISES

1. Assume that the probability of a randomly selected individual in a certain town being a Democrat is $\frac{1}{2}$. What is the probability that out of the first 6 people one meets, just 2 are Democrats? The probability that at least 2 are Democrats?

2. Assuming that vital statistics in the United States for the past twenty-five years show the constant probability of getting a boy to be equal to $\frac{2}{3}$ and the probability of getting a girl, $\frac{1}{3}$, what is the probability of a randomly selected mother having 2 boys out of 5 children?

3. Which of the following is true of the binomial distribution but not of the normal distribution: (a) Implies discrete data. (b) Is not always symmetric. (c) Has a known mean and standard deviation.

CHAPTER 5

SAMPLE DESIGN

Let us assume that 150 families in a small city are questioned about their annual income. It is found that the mean annual income of the 150 families is $3,000.

But to know the mean annual income of these 150 families may not be our sole interest. We may want to generalize from this random sample to a population or universe. We may want to estimate the mean annual family income for the city as a whole.

We are defining *population* or *universe* (the words are used synonymously) as the whole set of observations of some common characteristic about which we are interested in gaining information—in our example, the family income of every family in the city. A *sample* is a subset of these observations—here, family income of 150 city families. Note that although we speak about a universe of people or a sample of people, the units in the universe or sample may be not persons but observable characteristics like income or urban-rural residence or age.

The process of generalizing from a sample to its universe is an inductive process. We cannot guarantee the truth of the induction. There is no logical necessity which makes the random sample a representative one.

We can, however, determine the risk of error in our universe estimate, that is, the chance error due to the study of only a sample of the universe and not the universe itself. There is one stipulation. The risk of chance error can be determined only if the sample is a probability sample.

5.1. Probability Sampling

A probability sample is one drawn by a *probability method* of sampling. A probability method of sampling ensures every member of the universe a known, or determinable, chance of being drawn into

the sample. The kinds of probability samples we shall discuss are (1) simple random samples; (2) stratified random; (3) systematic; (4) cluster samples; and (5) a combination of stratified and cluster sampling.

Simple Random Sampling. A sample is considered a *simple random one* if its members are drawn in such a way that each observation of the universe has an equal chance of being included in the sample, and every possible combination of observations in the universe has the same chance of being included.

A first prerequisite for simple random sampling is some identification (by number, or its equivalent) for each observation in the universe. Then to ensure random selection of sample observations, eliminating all possibility of personal choice, such mechanisms as the following are used. All observations are identified on identical slips of paper that are put into a fish bowl, shuffled well, and a sample is drawn from this universe. Or a table of random numbers is used. If 150 families are desired in the sample out of a universe of 15,000, families are numbered from 1 to 15,000, and all numbers have 5 digits (e.g., 00001, 00002, . . . 15,000) so that each family has an equal chance of appearing in the sample. The first five-digit number is selected in a random manner from the table of random numbers, and then we continue to read five digit numbers in any consistent direction until 150 numbers between 1 and 15,000 are selected. Five-digit numbers higher than 15,000 (i.e., between 15,000 and 19,999) are skipped. If a number is drawn more than once, it is not included more than once in the sample.[1] Once sample members have been selected there is no substitution.

The results of random sampling may not look very random and may not be very typical or representative. For example, in tossing

[1] The stipulation that no family can be included more than once in the sample means that the probability of being selected is not constant. If, for example, we were to select a card at random from a deck of 52 cards, then return it to the deck, allowing it to be selected more than once, the probability at each try of any of the four suits being chosen is $\frac{13}{52}$. However, if the first card chosen, a spade, is removed from the deck so that it cannot be selected again, the probability of a spade being chosen on the second try is $\frac{12}{51}$. as compared with the greater probability ($\frac{13}{51}$) of a club, diamond, or heart being chosen. The selections are not independent; what was selected in the first try has some influence on what will be selected in the second. If the population is large relative to the size of the sample, the random sampling requirement that each member of the population have an equal chance of being selected is not seriously affected by sampling without replacement.

a perfect coin 20 times, we may get 20 successive heads. This is highly unlikely, but it is possible. As the sample becomes larger, it is more likely that the random sample will be representative of its universe. The mere inspection of a sample will not tell whether or not it is random; we must know how it was drawn.

Frequently, a complete listing or numbering of the universe is difficult or impossible to obtain. Even if there is such a listing, the time and expense involved in drawing and interviewing a random sample from this universe may be prohibitive. Futhermore, less reliable information may be obtained from simple random sampling than from other probability sampling procedures.

Stratified Random Sampling. The universe is divided into subgroups, called strata, on the basis of some characteristic related to family income, which is the variable being studied; a simple random sample is then taken from each stratum.

One basis of stratification might be age of family head, a relevant variable for which we have information. Family heads under 30 years of age are included in the first stratum, family heads between 30 and 40 years form the second stratum, and so on. Every family head is included in only one stratum. The strata are mutually exclusive and exhaustive. The strata boundaries are fixed so that family income is as *homogeneous as possible within any age stratum*, and as heterogeneous as possible between the age strata.

A sample of families is selected at random from each of the age strata, the total sample from all strata adding up to the desired sample of 150. How large a sample to select within each stratum? We may select the sample within each stratum proportional to the size of the stratum. If 20 per cent of the family heads in the universe are in the age stratum under 30, then 20 per cent of our sample comes from the first age stratum. The allocation of a sample to strata on a proportional-to-size basis is especially useful where we are stratifying to obtain a variety of information, for example, not only family income, but also political affiliation and employment status. When we select a sample for only one purpose—to determine family income—we would want to sample more intensively in those strata where income is heterogeneous. If all the family heads over 65 receive approximately the same income, then a very small sample of family heads of that age group would be adequate; but if the income received

by family heads under 30 is exceedingly variable, then we would sample more intensively in the under-30 age group.

Stratified random sampling leaves less to chance than does simple random sampling. If we have stratified well and allocated our sample well within the strata, we can get a more reliable estimate of a universe value from a stratified random than from a simple random sample.

Systematic Sampling. The first member of the sample is selected in a random manner, and then every *n*th unit is included in the sample, *n* being the quotient of size of universe divided by size of sample. If a sample of 150 families is desired out of a universe of 15,000, we can select every one hundredth family (15,000, the size of the universe, divided by 150, the size of the sample, equals 100), starting with some random number between 1 and 100. To select this random number, we number the first 100 families, and select a number between 1 and 100 from a table of random numbers. Beginning with this randomly selected family, we then take every one hundredth family.

Systematic sampling is suitable only if the periodic order of elements—every one hundredth family—has no relation to income, the characteristic being sampled. If every one hundredth family lives in a corner house where family income may be relatively higher than in noncorner houses, the sample would be biased accordingly.

Systematic sampling is often used where no periodic arrangement seems evident, since: (1) it is easier to select every *n*th family than to select a simple random sample; (2) where the universe is ordered in a continuous but not periodically recurring fashion that is relevant to the problem, e.g., with the lowest income group at the beginning and the highest at the end, systematic sampling is roughly equivalent to stratification and can be consequently more reliable than simple random sampling. If, out of 15,000 families, we select every one hundredth family in some prearranged order, and if the ecological pattern of the residential areas of the city roughly follows the socio-economic level, we are roughly stratifying by socio-economic strata, which means that we have a greater probability of getting a proper income representation than from a simple random sample.

Cluster Sampling. The sampling units selected are clusters of more than one element, e.g., city blocks of dwelling units. Cluster

sampling has the advantage of costing less per unit sampled than simple random sampling, because of the proximity of each member of the cluster; it has the disadvantage of not being as representative of the varied elements of the universe as a simple random sample insofar as there is close association among elements of a cluster. For example, all the dwelling units in homogeneous city blocks probably approximate each other more closely, and would be less representative of the whole city, than an equivalent number of dwelling units selected at random from the city.

A Combination of Stratified and Cluster Sampling

Problem. After World War II, rent controls were removed in areas when the supply of rental housing was considered sufficient to meet the effective demand. In the absence of controls, it is assumed that landlords will charge what the market will bear. Other conditions remaining relatively stable, rentals will presumably not increase after the removal of controls if the supply of rental housing is ample.

Assume that rent controls were removed in the city on December 31, 1950. We want to determine how much the average monthly rental increased for the year ending December 31, 1951.

Procedure. (1) We divide the city into strata according to characteristics that are significantly related to rent increases. A significant factor might be the rent level of the dwelling unit. For example, do those families who pay the lowest rent have the greatest percentage increase in rent? Another relevant factor might be the income of the tenant (or of the landlord), or the geographic area in which the dwelling unit lies. We shall assume that geographic area has a high association with income and with rent, and we shall stratify by geographic area.

The city is divided into six strata according to some determination of strata boundaries. Such natural or artificial boundaries as rivers, highways, and railroad tracks may be helpful in the determination of strata limits, especially when they demarcate socio-economic areas.

(2) On a map of the city, we mark off the six strata and list all the blocks within each stratum.

We then randomly select a given number of blocks from each of the strata. The blocks are clusters that help reduce our sampling cost. It is cheaper to concentrate interviews in a geographically small area than to have interviews dispersed throughout the city. Although

economical, cluster sampling tends to increase sampling error insofar as the units within a cluster are alike, any single cluster thus giving less information about the universe than an equivalent number of randomly selected units.

(3) Next, interviewers list all the rented dwelling units in the blocks selected, starting at one end of the block and going around the block in a prearranged direction. This is another economy over simple random sampling. We need list only those rented dwelling units in the randomly selected blocks within each stratum; we need not list all the rented dwelling units in the city as a whole.

Finally, we select at random a sample of rented dwelling units within each sample block.

(4) We now have our sample. We shall not discuss here the problems that arise, such as: (a) Have the present tenants been in the dwelling units since the beginning of the year; do they know the rent rate since the beginning of the year? (b) Are there chronic refusals or absentees, even after the interviewer returns a different day and a different hour from previous visits; do these self-selected nonrespondents differ in rent increases incurred from the rest of the sample?

5.2. What Kind of Errors Can Be Made in Generalizing from a Sample to a Universe?

If family income were highly homogeneous, that is, if all families had almost the same income, then we could sample only a few families to determine income. Taking a sample of only a few drops of blood, for example, assumes great homogeneity in a person's blood, and the error in generalizing from the sample should be small. Most sociological data are not so homogeneous.

The kinds of error possible in generalizing from a sample to a universe can be classified into two categories: (1) errors of chance and (2) errors of bias.

(1) *Errors of Chance.* In *generalizing* from a sample, error is possible because we are looking at just a part and not the whole of the universe.

For example, if our universe consists of 10 families, and we want to select a sample of 5 of these 10 families to determine mean family income, there are many different combinations, or different ways in which 5 families can be selected out of 10. The mean family income

of the samples of 5 families would differ among themselves and would differ from the universe mean:

THE UNIVERSE

Family	*Annual Income*
1	$3,900
2	2,700
3	2,800
4	3,200
5	2,400
6	2,500
7	3,800
8	3,200
9	2,400
10	2,900

The mean family income of the universe is $2980. The mean income of sample of families 1, 2, 3, 4, and 5 is $3000. The mean income of sample of families 2, 4, 6, 7, and 10 is $3020, etc.

If we select 150 Ann Arbor families out of a universe of 15,000 families, the number of possible samples is of course very much greater than the selection of 5 families out of 10.

(2) *Errors of Bias.* In addition to errors due to chance fluctuations in sampling, there are other errors due to bias. *Bias* exists if repeated samples differ systematically, on the average, from their universe in the characteristic studied. Some observations in the universe have a consistently greater probability of being included in the sample than do others. In unbiased sampling, the probability of greater accuracy increases as the size of the sample increases. This is not true in biased sampling.

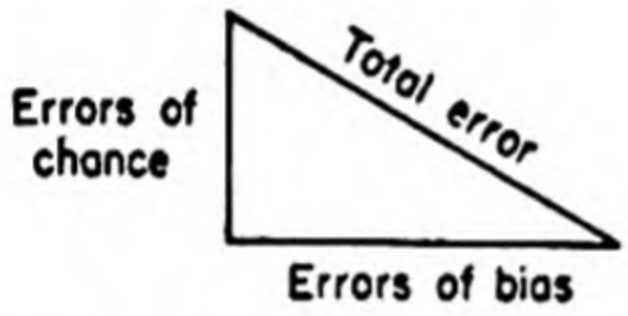
Errors of bias due to faulty observation or measurement can occur in any interview situation, whether a sample or a complete universe. For example: (a) The question may have different meanings according to the social norms of the respondents. (b) The respondent may give the answer that he thinks the interviewer wants rather than the one he believes to be true. He may answer incorrectly in resentment against the interviewer or in distrust of the avowed anonymity of the interview. He may forget the correct answer. (c) The interviewer, with his own personal biases, may distort the respondent's answer.

In addition, errors of bias can arise in the sampling process. The initial selection of respondents may not be representative of the universe to which we want to generalize. Alternatively, those who re-

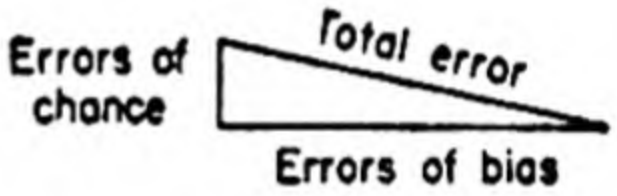
fuse, or are unable to be interviewed, may differ in some crucial way from the respondents. If the not-at-homes differ in the characteristic studied from those who stay at home and if the refusals differ from those willing to be interviewed, then nonresponse introduces an error of bias.

If we want to visualize errors geometrically, we can think of total error as the hypotenuse of a right triangle, with errors of chance and errors of bias being the two legs of the triangle.[2] The diagram illustrates that to reduce total error both errors of chance and errors of bias must be reduced. When errors of bias are large, it is ineffectual to get a large sample in order to keep chance errors small.

Relative Size of Initial Errors



Reducing Only Errors of Chance Does Not Greatly Reduce Total Error



Errors of chance occur only in sampling. Errors of bias can occur in a universe study as well as a sample study. Since errors of bias can exceed errors of chance, a sample estimate may conceivably be more accurate than a universe estimate. For a given sum of money, concentrating on a sample may produce better trained interviewing personnel and more uniform methods than a complete study of the universe. A sample has advantages of economy in time and money.

In probability sampling—simple random or modified random—the probability of selection is known for each member of the universe. It is only in probability sampling that statistical theory can be applied to determine the risk of chance error. There is, however, a major difficulty to be considered in applying probability sampling to actual surveys. Probability sampling is expensive in time and money where many call-backs are necessary to try to interview a certain respondent in the household. It is somewhat easier if any adult in the household can respond. Where there is a sizable non-respondent or refusal rate, the meaningfulness of calculating the risk of chance error based on probability theory is questionable.

[2] Deming, William Edwards, *Some Theory of Sampling* (New York: John Wiley & Sons, Inc., 1950), pp. 25–26 and 129–30.

5.3. Quota Sampling

A quota sample is a nonprobability sample. In quota sampling there is a personal choice in the determination of sample members.

A sample may be controlled through subdividing the universe by those variables that seem to have a high degree of relationship with the characteristic being studied. Quota sampling uses such controls; it differs from what we have called "stratified random sampling" in that it selects its respondents within each stratum in a nonrandom manner.

Gallup, Roper, and Crossley have employed quota sampling in their public opinion polls. To survey presidential election preferences, they divided the United States into areas, the proportion of the total sample in each area being determined by such characteristics as the number in the area eligible to vote and the number voting in earlier elections. Within each area, interviewers were assigned quotas based on the distribution of characteristics considered relevant to voting, such as economic status, party affiliation, sex, and age. An interviewer, for example, was told to get a quota of 50 males and 40 females in a certain area, possibly in a certain age range and in a specified income group, who voted Democrat in the last election.

By this method of sampling, the interviewer is free to choose the particular respondents he will interview, subject to the condition that he pigeonhole his respondents into his quota. Bias may arise because the interviewer chooses the most easily contacted respondents. He may choose people who are outside rather than inside their homes, or he may overrepresent people who live on the first floor rather than the tenth, and who are in the middle income rather than the highest or lowest income brackets (if income is not controlled).[3] He may look for people who agree with his point of view. Attempts may be made to control such biases.

[3] In their comparative study of probability and quota sampling on the adult population of Iowa, Haner and Meier found that the probability samples had more households in the upper socio-economic brackets than did the quota samples. This finding is at variance with the popularly held assumption that upper-income persons are oversampled in quota samples, when income is not controlled. Haner and Meier tried to keep all conditions constant except *method* (the same schedules were used, the same interviewers, and the same universe) in comparing quota and probability sampling. C. F. Haner and N. C. Meier, "The Adaptability of Area-Probability Sampling to Public Opinion Measurement," *The Public Opinion Quarterly*, Summer 1951, pp. 335–52.

Quota sampling eliminates the necessity for call-backs to get specific respondents. The chief difficulties in the use of quota sampling are that (1) the variables related to the characteristic being studied may not be known; (2) even if these relevant variables are known, their distribution in the population may not be known. These two difficulties are encountered in any kind of stratified sampling. In addition, (3) bias can be introduced into the selection of respondents by the interviewer. The risk of chance error cannot be estimated since the samples are not selected in a random manner.

KEY TERMS

cluster sampling
errors of bias
errors of chance
population
probability sampling
quota sampling
sample
simple random sampling
stratified random sampling
systematic sampling

REFERENCES

Hagood, M. J., and D. O. Price, *Statistics for Sociologists*, chap. 18. New York: Henry Holt and Company, 1952.

Parten, Mildred, *Surveys, Polls and Samples: Practical Procedures*. New York: Harper and Bros., 1950.

Peatman, John G., *Descriptive and Sampling Statistics*, chap. 11. New York: Harper and Bros., 1947.

EXERCISES

1. We want to determine the proportion of students in a statistics class who intend to go into the teaching profession. We do not want to question each person in the class of 80, but only a random sample of 20 students.
   (a) What is the first step in selecting a random sample?
   (b) Use the table of random numbers to select a random sample of 20.
   (c) What kinds of errors might appear in the interview results?
   (d) If you were using a quota instead of a probability sample, what would be your procedure? What possible errors would you have?

2. If you want to find out from a sample of students the proportion of students at the university smoking various brands of cigarettes:
   (a) How would you select a sample?
   (b) Would you enumerate all possible brands in your presentation of results? How might the data be condensed?
   (c) What kinds of errors might the results contain?

3. If you want to find out what per cent of working sons in a certain city are in the same occupational stratum as their fathers, what per cent are in a lower, and what per cent, a higher stratum:

(a) How would you select a 1,000-son sample?

(b) How would the data be tabulated?

(To determine strata, you can use the occupational hierarchy devised by Alba Edwards, Bureau of the Census.)

(c) What kinds of errors might you have?

CHAPTER 6

INTRODUCTION TO STATISTICAL INFERENCE

We want to estimate the value of a universe *parameter* (e.g., a universe mean, or any other measure of an entire population) from a sample *statistic* (a sample mean, or any other measure based on a sample). The universe is a set of all possible observations or measurements. The sample is a part of this set. The universe mean and standard deviation are symbolized by M and σ, respectively; the sample mean and standard deviation, by $\overline{X}$ and s.

Inferring from a sample to a universe always involves uncertainty. Different random samples of the same size from the same universe will give different estimates of the universe mean. If the sample has been selected from its universe in a random manner, we can determine rather precisely the risk of chance error in our inference from the sample statistic to the universe parameter. Our confidence or lack of confidence in the inference can be expressed in probability terms: "We are 95 per cent confident in our estimate," or "We are 99 per cent confident."

To understand the basis for being able to determine chance error, we must first understand the concept of a sampling distribution.

6.1. Sampling Distributions

Assume that a hypothetical universe consists of four boys, ages 2, 3, 5, and 6. The mean age of the universe of four boys is 4, the variance, $\frac{10}{4}$, or 2.5.

We want to select samples of 2 from the universe of 4 boys.[1] We write the 4 ages on 4 different slips of paper, put the papers into a fish

[1] In statistical application, the universe will always be larger than 4, the sample larger than 2. Small figures are used here for illustrative purposes.

Table 6-1. Computation of Mean Age and Variance for Universe of Four Boys

Boy	*Age* (X)	*Deviations from Mean* (x)	*Squares of Deviations from Mean* (x^2)
1	3	−1	1
2	2	−2	4
3	5	1	1
4	6	2	4
Sum:	16		10

Universe Mean: $M = \frac{\Sigma X}{N} = 4$ Universe Variance: $\sigma^2 = \frac{\Sigma x^2}{N} = \frac{10}{4} = 2.5$

bowl, and draw all possible samples of 2 boys. There are 16 different possible samples, assuming that (1) any boy can be chosen more than once in the same sample—we return each slip of paper to the fish bowl as soon as the age is recorded—and that (2) if the same 2 boys are drawn in different order in 2 samples, we include both samples. The 16 possible samples of size 2 out of a universe of 4 boys are:[2]

	Age		*Age*		*Age*		*Age*
(1)	2,2	(5)	3,2	(9)	5,2	(13)	6,2
(2)	2,3	(6)	3,3	(10)	5,3	(14)	6,3
(3)	2,5	(7)	3,5	(11)	5,5	(15)	6,5
(4)	2,6	(8)	3,6	(12)	5,6	(16)	6,6

If we compute the mean of each of the possible samples of size 2 from the universe of four boys (Fig. 6.1), the sample means will form a distribution (Table 6-2). This distribution is called the *sampling distribution of means*. We see from Table 6-2 that the mean of all possible sample means is 4, which is the mean of the universe.

[2] If we were to disregard the order of selection, and we did not replace the slips of paper into the fish bowl after each drawing, there would be only 6 different possible samples of size 2 from a universe of 4. The samples are:

	Age		*Age*
(1)	2,3	(4)	3,5
(2)	2,5	(5)	3,6
(3)	2,6	(6)	5,6

The number of such combinations can be determined by the formula

$$C_X^N = \frac{N!}{X!\,(N-X)!}$$

where $X = 2$ and $N = 4$ $\left(C_2^4 = \frac{4!}{2!\,2!} = 6\right)$

FIRST SELECTION

SECOND SELECTION Age	Age 2	3	5	6
2	2	2.5	3.5	4
3	2.5	3	4	4.5
5	3.5	4	5	5.5
6	4	4.5	5.5	6

Fig. 6.1. The mean age of all possible samples of size 2 from a universe of 4 boys, ages 2, 3, 5, and 6. (The ages in the margin give the sample selected; the intersection of the two selections gives the mean of each sample.)

Table 6-2. Sampling Distribution of Mean Age
(Samples of size 2 from universe of 4 boys)

Mean Age ($\bar{X}$)	*Frequency* (f)	*Frequency Times Mean Age*
2	1	2
2.5	2	5
3	1	3
3.5	2	7
4	4	16
4.5	2	9
5	1	5
5.5	2	11
6	1	6
Sum:	16 ($= n$)	64

The Mean of All Possible Sample Means Is Equal to the Mean of the Universe: $\frac{\Sigma f\bar{X}}{n} = 4$

We next compute the variance of the sampling distribution of means. (You will recall that the variance is the square of the standard deviation.)

Note in Table 6-3 that the variance of the sampling distribution of means (1.25) is equal to the variance of the universe (2.5) divided by n (the size of each sample, equal to 2). Consequently, the standard deviation of the sampling distribution of means is equal to the standard deviation of the universe divided by the square root of n. The standard deviation of the sampling distribution of means is called the *standard error*. It is written $\sigma_{\bar{X}}$.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \tag{14}$$

where σ = the standard deviation of the universe and n = the sample size.

Table 6-3. The Variance of the Sampling Distribution of Mean Age
(Samples of size 2 from universe of 4 boys)

Mean Age $\overline{X}$	f	$f\overline{X}$	x	x^2	fx^2
2	1	2	−2.0	4.00	4.0
2.5	2	5	−1.5	2.25	4.5
3	1	3	−1.0	1.00	1.0
3.5	2	7	−0.5	0.25	0.5
4	4	16	0	0.00	0.0
4.5	2	9	0.5	0.25	0.5
5	1	5	1.0	1.00	1.0
5.5	2	11	1.5	2.25	4.5
6	1	6	2.0	4.00	4.0
Sum:	16	64			$\Sigma fx^2 = \underline{20.0}$

$$\frac{\Sigma fx^2}{n} = \frac{20}{16} = 1.25 \qquad \left(= \frac{\sigma^2}{n} = \frac{2.5}{2}\right)$$

The Variance of the Sampling Distribution of Means
Is Equal to the Variance of the Universe Divided by n

The histogram for our sampling distribution of means is given in Fig. 6.2.

Fig. 6.2. Histogram for sampling distribution of mean of all possible samples of size 2 from a universe of 4 boys.

Whatever the form of the distribution of the universe, *the sampling distribution of means approximates a normal curve if the sample size is 30 or larger.*

This is very useful information. It tells us that if we draw a single random sample of size 30 or larger from its universe, and compute the sample mean, this sample mean will fall somewhere along a normal curve whose mean is the mean of the universe, and whose standard error is the standard deviation of the universe divided by the square root of n.

We can then say, for example, that only 5 per cent of the sample

means will differ from the mean of the universe by more than 1.96 standard errors, and that only one per cent of the sample means will differ from the mean of the universe by more than 2.58 standard errors. (See Normal Curve Table, page 168.)

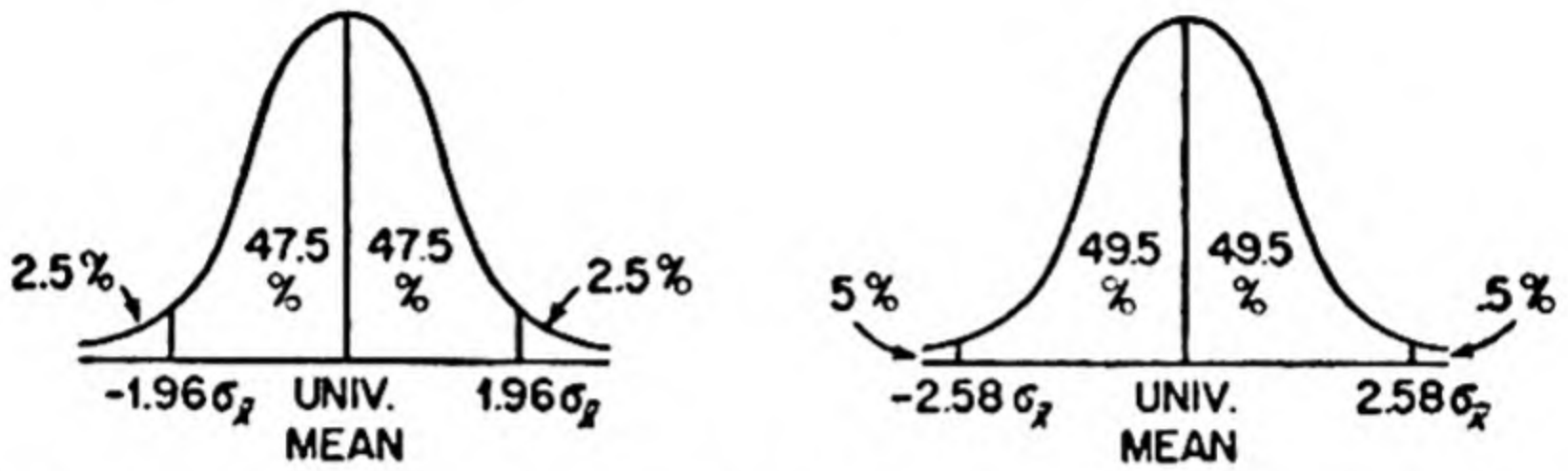


Fig. 6.3. Sampling distribution of means forms a normal curve.

We do not know whether or not our sample mean is in the tail of the normal curve. But if only 1 out of every 20 sample means is in the tail, the probability is only 5 per cent that our sample mean is in the tail of the curve. Alternatively, if only 1 out of every 100 sample means is in the tail, the probability is 1 per cent that our sample mean is in the tail of the normal curve.

Note that the standard error ($\sigma/\sqrt{n}$) becomes smaller as we increase the size of the sample. If n equals 36, then the standard error equals $\frac{1}{6}\sigma$; if n equals 100, the standard error equals $\frac{1}{10}\sigma$. As the sample size increases, the sampling distribution of means clusters more closely about the universe mean.

In practice, we do not actually determine the distribution of all possible random samples of the same size in a universe. We usually select only one or possibly two samples from a universe. It is important, however, to have an intuitive understanding of the concept of a sampling distribution.

6.2. Two Methods of Statistical Inference

There are two basic methods of statistical inference: (1) The estimation of a universe parameter (e.g., a universe mean or proportion) from a sample statistic (a sample mean or proportion); (2) The testing of hypotheses about the universe parameter.

Estimating a Universe Parameter. The mean of the universe is unknown. We know the mean of a single sample.

We know too that means from all possible random samples of the same size are normally, or approximately normally, distributed around the universe mean if the sample is large or the variable normally distributed in the universe. Consequently, we can set up *confidence limits* on both sides of the sample mean and estimate, with a certain risk of error, that the universe mean lies somewhere in the interval within these limits. We might set our confidence limits at $\bar{X} \pm 1.96$ standard errors. We would then say that our confidence interval has a 95 per cent chance of containing the mean of the universe, since 95 per cent of the intervals obtained in this manner will include the universe mean. We say we are 95 per cent sure that our interval contains the universe mean; we have odds 19 to 1. These are the odds usually demanded by a social scientist.

Testing Statistical Hypotheses. We set up statistical hypotheses that include (1) a null hypothesis and (2) alternative hypotheses. The *null hypothesis* states that the universe mean (or some other parameter) is equal to a certain value. The *alternative hypotheses* state that the universe mean is different from this null hypothetical value, e.g., that it is greater than this hypothetical value, or that it is less than this value.

We perform a statistical test to decide whether to accept the value of the mean stated in the null hypothesis or in the alternative hypothesis. In determining the choice of the test, we want to minimize the two possible types of error: the error of rejecting the null hypothesis when it is true and the error of accepting the null hypothesis when it is false. A good test is one that makes the probability of both errors as small as possible.

To test our hypotheses, we shall select only one sample from the great number of possible samples in our universe. We have noted before that, if the sample is sufficiently large, the sampling distribution of means of all possible random samples of the same size from a given universe forms an approximately normal curve, with the mean of the sampling distribution of means equal to the mean of the universe, and its standard error equal to the standard deviation of the universe divided by the square root of n. Consequently, we can measure the area under the normal curve between our *sample mean* and the *universe mean* of the null hypothesis and we can say how probable it is that our sample mean comes from this hypothetical universe. We know that the normal curve reaches out infinitely far

in both directions, so there is some probability of getting a sample mean very far removed from its universe mean, but the chance is exceedingly slight.

If the distribution of sample means follows the normal curve, there will be a proportionately greater number of sample means for a given standard-error space close to the universe mean, and the proportion will get successively smaller as we go out to the tails of the distribution. We arbitrarily set up a region of acceptance and say that if a sample mean is within this region, then the difference between the sample and universe mean is due purely to chance, and our null hypothesis cannot be rejected. If the sample mean is outside this region, then we say that the difference between the sample and the universe mean is not due to chance, but is significant, and we reject the null hypothesis and accept the alternative hypothesis. The limit of the region at which we say we shall no longer regard chance as operating, and hence must reject the null hypothesis, is our level of significance.

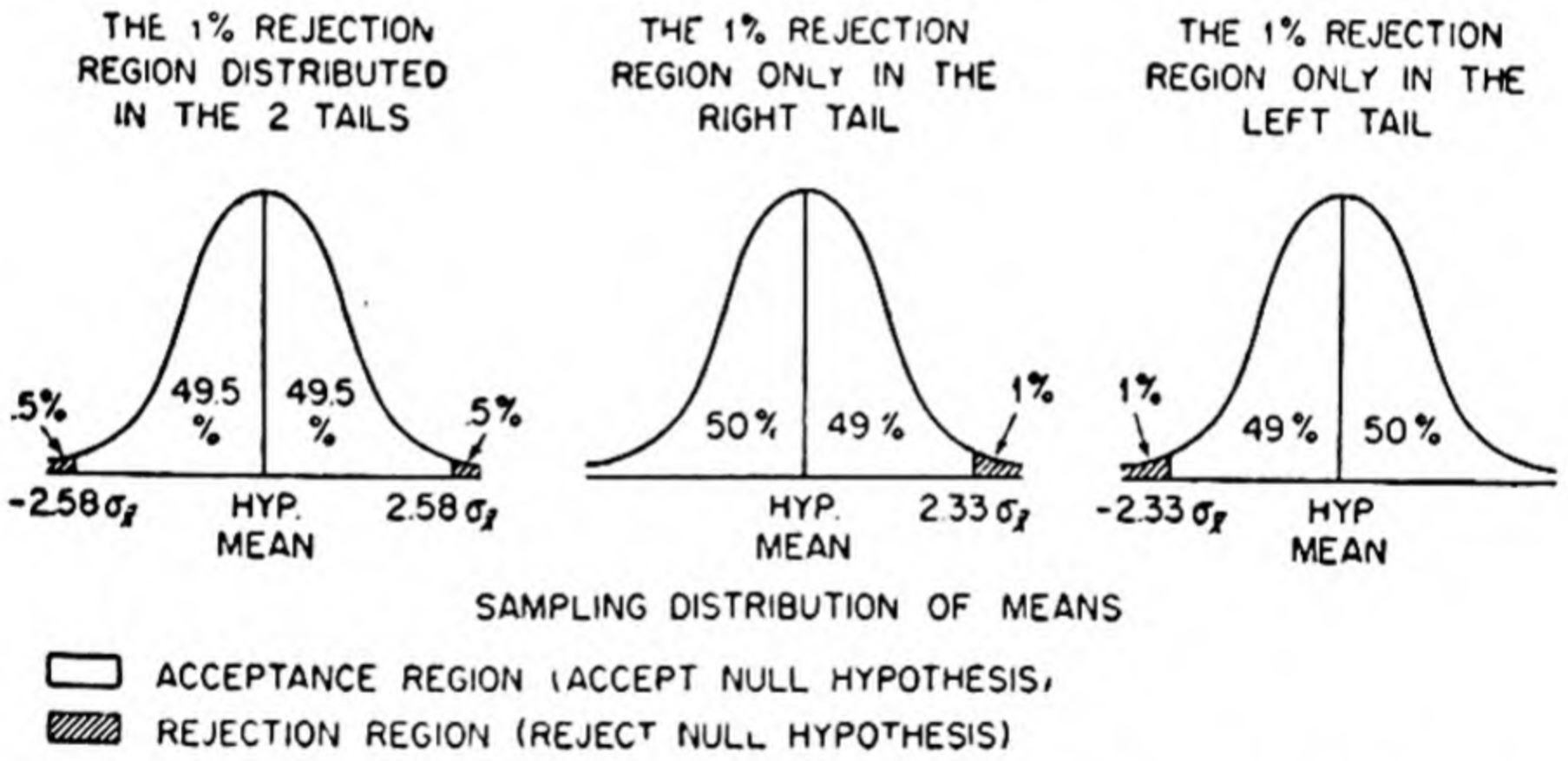


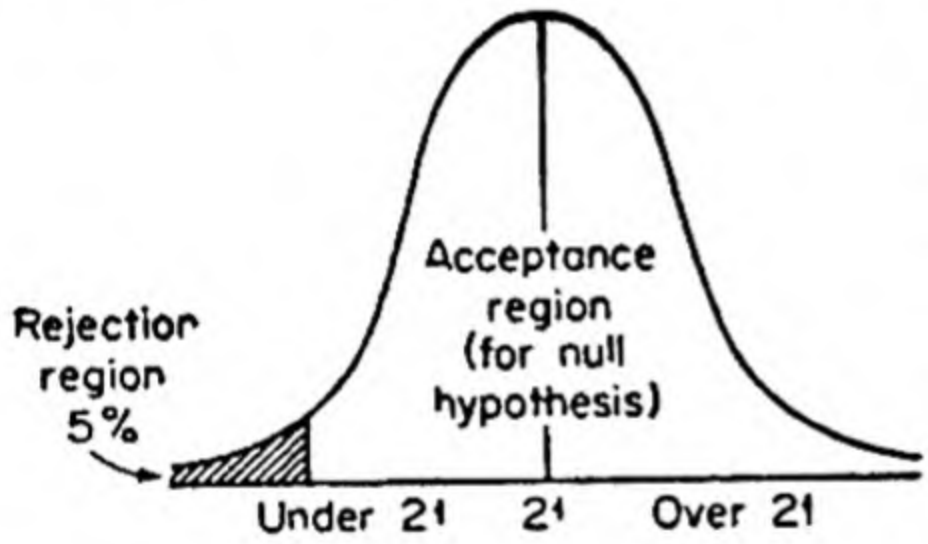
Fig. 6.4. Rejecting the null hypothesis at the 1 per cent level of significance.

Where do we set our significance level, which divides the acceptance region from the rejection region? Since we shall examine only a sample, a part of the universe, and not the complete universe, we can never be sure that we have not made an *error in rejecting the null hypothesis, when it actually should have been accepted.* This is an error of *wrong rejection* (commonly referred to as a Type I error). The risk of this type of error can be made as small as we like by setting

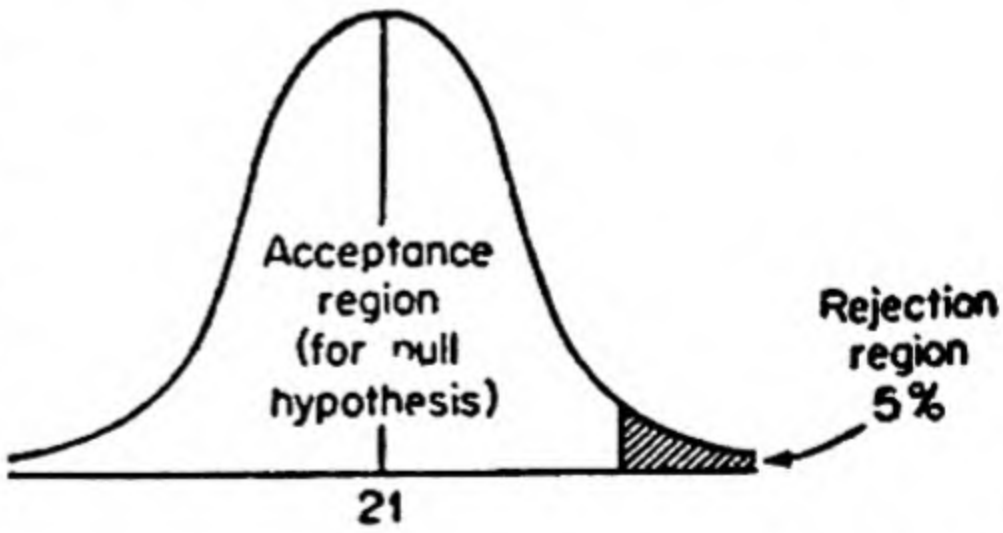
the significance level far out in the tail of the normal curve. Instead of a 5 per cent rejection region, we can have a 2 per cent or 1 per cent or .5 per cent region. In most statistical studies in the social sciences, the level of significance is set at 5 per cent, 1 per cent, or .5 per cent, and the rejection region is distributed in either one or both tails of the normal curve.

But the farther out in the tail we set our level of significance, the more the risk of another type of error: we can *accept a null hypothesis when it should have been rejected.* This is an error of *wrong acceptance* (referred to as a Type II error).

The probability of accepting a false null hypothesis depends upon the actual value of the universe mean. If the null hypothesis value is close to the actual value, then we can easily make the mistake. If the value is a great distance away from the actual universe value, it is less likely that we will accept a false null hypothesis.



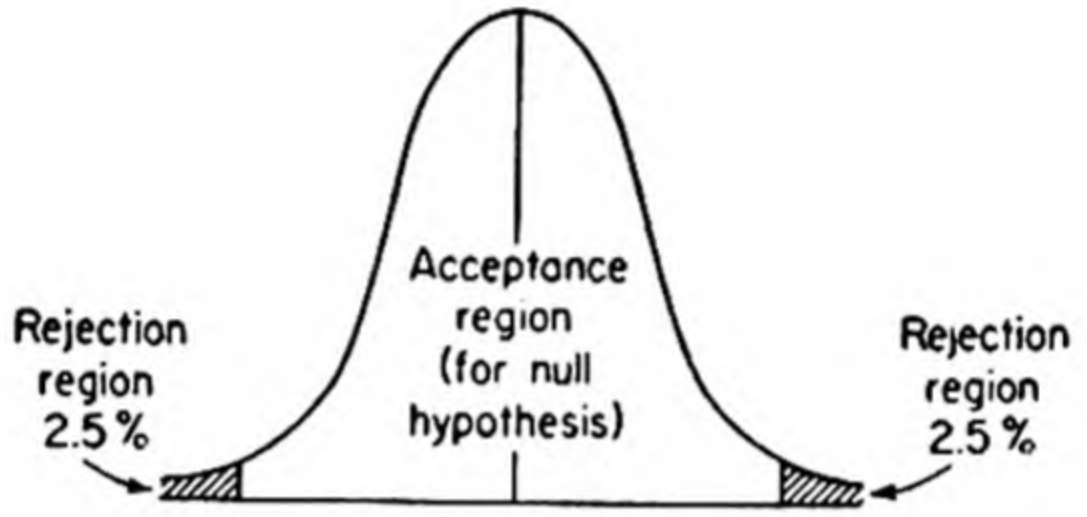
We can minimize the risk of accepting a false hypothesis by placing our rejection region in that area of the curve where lie the alternative values against which we are interested in testing our null hypothesis. For example, we may want to test the null hypothesis that the average age of college seniors is 21 years. Hence if age is approximately normally distributed among college seniors, about one-half of them can vote in the national elections. Our alternative hypotheses state that the mean age of students is less than 21, and that more than one-half the students will thus be affected by possible reduction in the age of voting. The alternative we are interested in testing against is a mean age *less than 21.* Therefore we put our 5 per cent rejection region in the left tail of the normal curve. By doing this, we are increasing the power of the test to reject our 21-year hypothetical mean in



favor of our alternative. The *power* of a test is the probability of rejecting a null hypothesis when it is false.

If the alternatives, against which we are interested in testing the null hypothesis, are *higher than* 21, we put the 5 per cent region of rejection in the right tail of the normal curve. We are again increasing the power of the test to reject our 21-year hypothetical mean in favor of the alternative.

If our alternative hypotheses state that the mean age of Seniors may be higher or lower than 21, we distribute the 5 per cent rejection region in the two tails of the normal curve.



KEY TERMS

acceptance region
alternative hypotheses
confidence limits
level of significance
null hypothesis
parameter
power of a test
rejection region
sampling distribution
standard error
statistic
two types of error

REFERENCES

Dixon, W. J., and F. J. Massey, *Introduction to Statistical Analysis*, chaps. 4 and 7. New York: McGraw-Hill Book Company, 1951.

CHAPTER 7

STATISTICAL INFERENCE CONTINUED

7.1. The Mean: Testing Hypotheses and Making Estimations

Problem. Studies have been made of the social and psychological advantages to children of foster-home over institutional living. In foster homes, children presumably can get the parental and family love that an institution cannot give.

We want to find out whether foster-home living has any effect on the intelligence level of the child.

Null Hypothesis. We set up a null hypothesis that, on the average, there is no change over a three-year period in the intelligence quotient of children transferred from institutions to foster homes.

Our null hypothesis of no change in IQ is only a straw man, and we do not necessarily believe it to be true. We want to test this null hypothesis against the alternative hypotheses that foster-home living will have a positive effect on the IQ of foster children. Consequently, to increase the power of rejecting the no-change in IQ hypothesis if it is false, we shall put our region of rejection only in the right tail of the normal curve.

We set up our level of significance to determine at what point the hypothetical universe mean of no change in IQ will be rejected, because the mean of our sample appears too far above the hypothetical universe mean to consider it a sample from that universe. We could conceivably get, strictly by chance, a sample mean that is a very great distance from the universe mean, but this would happen only rarely. If we set our level of significance at 5 per cent, then we will reject our null hypothesis where the positive difference between sample mean and universe mean could have occurred in only 5 random samples out of 100 from this universe. We regard the 5 per cent area a statistically significant region, the region of rejection of the null hypothesis in favor of the alternative hypotheses. The other 95 per cent of the area under the normal curve is a chance

region. If the difference between the sample mean and the hypothetical universe mean lies in the 95 per cent portion, it is considered a chance difference due to fluctuations of random sampling.

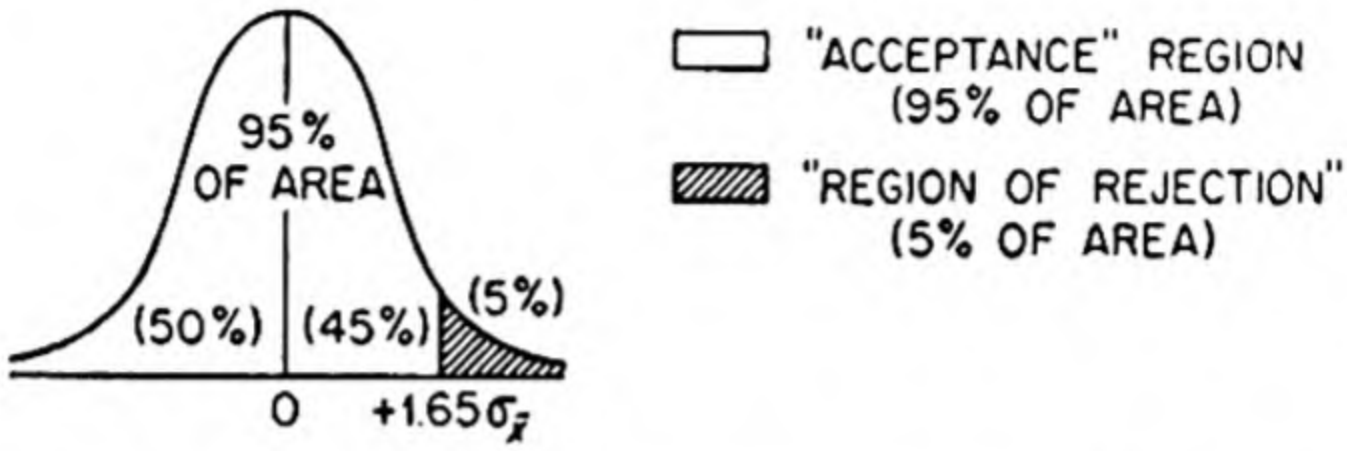


Fig. 7.1. Distribution of sample means around universe mean, 5 per cent significance level. (One-tailed alternative.)

A *z* score (the familiar standard score) of +1.65 standard errors is associated with a 5 per cent significance level in the right tail of the curve. If the positive difference between sample mean and hypothetical universe mean is greater than +1.65 standard units, we say that it is unlikely that the sample could have come from a universe where the mean change in IQ is zero, since only 5 per cent of its sampling distribution is in this rejection region. Such occurrences could happen in only 5 out of 100 random samples if the null hypothesis were true. (See the Normal Curve Table, page 168, for the correspondence between distance from the mean of +1.65 standard errors and area under the normal curve of 45 per cent.)

If we were to set our significance level not at 5 per cent but at 1 per cent, a *z* score of +2.33 standard errors would be associated with a 1 per cent level in the right tail of the normal curve. Beyond +2.33

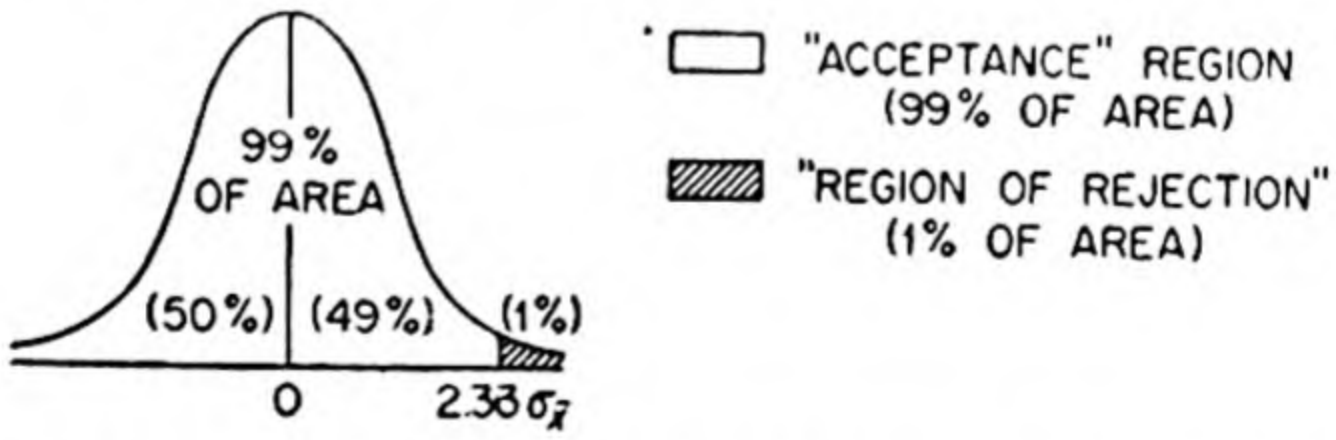


Fig. 7.2. Distribution of sample means around universe mean, 1 per cent significance level. (One-tailed alternative.)

standard units we reject the null hypothesis in favor of the alternative hypotheses, since only 1 per cent of the random sampling distribution

from the universe with the hypothetical mean is in this rejection region.

Procedure. The average gain over a three-year period in the IQ of 50 foster children between the ages of 4 and 6 is found to be 4 points, with a standard deviation of 6.7. These 50 children are a random sample of a universe of 1,000 children given out to foster homes by a certain institution from one to six months before the study. Table 7-1 gives the computation of the mean change in IQ and the standard deviation for the 50 foster children over the three-year period.

Table 7-1. Computation of Mean and Standard Deviation for Change in IQ of Fifty Foster Children Over a Three-Year Period
(Hypothetical data)

Change in IQ (X)	*Frequency* (f)	(fX)	*Deviations from Mean* (x)	(fx)	*Squares of Deviations from Mean* $f(x)$
−13	1	−13	−17	−17	289
−12	1	−12	−16	−16	256
−9	1	−9	−13	−13	169
−7	1	−7	−11	−11	121
−6	1	−6	−10	−10	100
−4	1	−4	−8	−8	64
−2	2	−4	−6	−12	72
−1	1	−1	−5	−5	25
0	3	0	−4	−12	48
1	2	2	−3	−6	18
2	3	6	−2	−6	12
3	5	15	−1	−5	5
4	6	24	0	0	0
5	5	25	1	5	5
6	4	24	2	8	16
8	3	24	4	12	48
10	2	20	6	12	72
12	2	24	8	16	128
14	2	28	10	20	200
15	2	30	11	22	242
17	2	34	13	26	338
Sum:	50 ($= n$)	200		0	2,228

Mean: $\frac{\Sigma fX}{n} = \frac{200}{50} = 4$

Standard Deviation: $s = \sqrt{\frac{\Sigma f(x)^2}{n}}$

$= \sqrt{\frac{2228}{50}} = \sqrt{44.56}$

$\doteq 6.7$

Does this sample give evidence of a statistically significant change in IQ in the three-year period? Can we rejectour hypothesis of no-change in IQ in favor of the alternative hypothesis of a positive change?

The universe mean gain in IQ according to the null hypotheses is 0, whereas the sample mean gain is 4. We want to translate this difference into standard error units, to determine how far away the sample mean is from the hypothetical universe mean in standard units. We use the familiar standard score, or z score, which is equal to the difference between the mean of the sample and the mean of the universe divided by the standard error of the mean.

$$\text{Standard Score: } z = \frac{\bar{X} - M}{\sigma_{\bar{X}}} \tag{15}$$

where $\bar{X}$ = mean of sample
M = mean of universe (according to the null hypothesis)
$\sigma_{\bar{X}}$ = standard error of the mean

We have said that the distribution of means from random samples of the same size in a universe is normal, or approximately so, if (1) the variable (here, IQ) is normally distributed in the universe, or (2) the sample size is at least 30. The mean of the distribution of sample means is the mean of the universe, and its standard deviation is the standard deviation of the universe divided by the square root of n. (The symbol n is the size of the sample.)
Standard error of the mean:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

σ = standard deviation of the universe
n = size of the sample

We do not know the standard deviation of the universe σ. If the sample is sufficiently large, the standard deviation of the sample s may be an adequate estimate of the universe standard deviation. The *estimated* standard error of the mean then becomes:

$$\text{est } \sigma_{\bar{X}} = \frac{s}{\sqrt{n-1}} \tag{16}$$

where s = standard deviation of the sample.

The standard deviation of our sample is 6.7. Consequently, the standard error of the mean can be estimated at .96.

$$\text{est } \sigma_{\bar{X}} = \frac{6.7}{\sqrt{50-1}} = \frac{6.7}{7} \doteq .96.$$

The mean of our sample is 4, the hypothetical universe mean is 0. The difference between sample and hypothetical universe mean is converted into standard error units.

$$z = \frac{\bar{X} - M}{\sigma_{\bar{X}}} = \frac{4-0}{.96} = \frac{4}{.96} = 4.2 \text{ standard error units}$$

A sample mean of 4 is +4.2 standard errors away from the hypothetical universe mean and could have happened by chance in less than 3 samples out of 100,000.

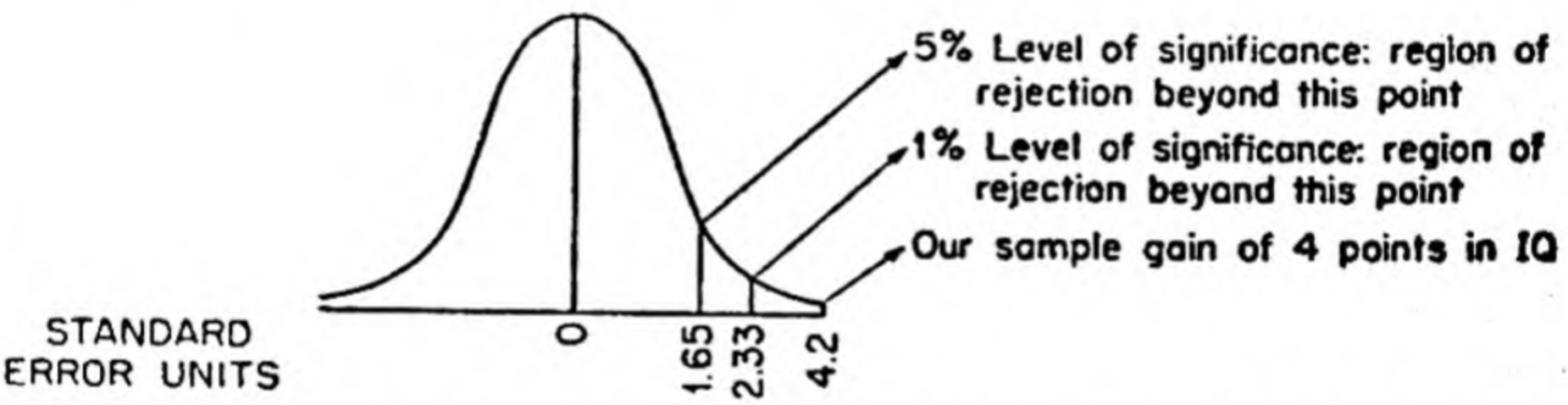


Fig. 7.3. Normal-curve distribution of mean change in IQ for all possible samples of size 50 from a universe with no mean change.

We reject the null hypothesis of no significant change in IQ over the three-year period and accept the alternative hypotheses that this sample gives evidence of an increase in IQ among foster children. We would reject the null hypothesis at both the 5 per cent and the 1 per cent level of significance. (In practice, we would set up one, and not two levels of significance, and this is done before the data are collected.)

Let us review the procedure in testing a statistical hypothesis:

(1) We state our null hypothesis: there is no change in the IQ of foster children over a three-year period.

(2) We state our alternative hypotheses: there has been an increase in the IQ of foster children over a three-year period.

(3) The significance level is set at 5 per cent. We reject our null

hypothesis if the difference between sample mean change and hypothetical universe mean change could have occurred in only 5 per cent (or less) of the random samples from this universe.

(4) The 5 per cent rejection region is put in the right tail of the curve. We place our rejection region in the area of the curve where lie the alternative values against which we are interested in testing our null hypothesis in order to minimize the risk of accepting a false null hypothesis.

(5) The acceptance region and the rejection region of the null hypothesis are determined. The region under the normal curve beyond +1.65 standard errors is considered the rejection region of the null hypothesis. If the difference between sample mean and hypothetical universe mean is less than +1.65 standard errors, we consider the difference due to chance errors of random sampling and not discrediting the null hypothesis. If the difference is greater than +1.65 standard units, we accept the alternative hypotheses.

(6) The *z*-score is computed from our sample:

$$\left(z = \frac{\overline{X} - M}{s\sqrt{n-1}}\right)$$

This gives us the approximate standard-error distance under the normal curve between the sample mean and the hypothetical universe mean.

(7) Either the null hypothesis or the alternative hypotheses are accepted according to whether the standard-error distance under the normal curve between sample mean and hypothetical mean is in the acceptance region (less than +1.65 standard errors) or the rejection region of the null hypothesis (+1.65 or more standard errors).

Making Estimations from a Sample Mean to a Universe Mean. We can use our sample mean to estimate the interval that contains the universe mean.

The sample mean gain in IQ is 4 points. This mean gain may be higher or lower than the mean change in IQ in the universe of foster children. We want to estimate the mean change in the universe with 95 per cent confidence.

We know that means from all possible random samples of size 50 (our sample size) are normally distributed around their universe mean. If the interval $M \pm 1.96\sigma_{\overline{X}}$ will contain 95 per cent of the sample means, then the interval $\overline{X} \pm 1.96\sigma_{\overline{X}}$ will contain the universe

mean in 95 per cent of the samples. (See Figs. 7.4a and 7.4b.)

The *confidence limits* are $(\bar{X} - 1.96\sigma_{\bar{X}})$ and $(\bar{X} + 1.96\sigma_{\bar{X}})$. The interval between $(\bar{X} - 1.96\sigma_{\bar{X}})$ and $(\bar{X} + 1.96\sigma_{\bar{X}})$ is called the *95 per*

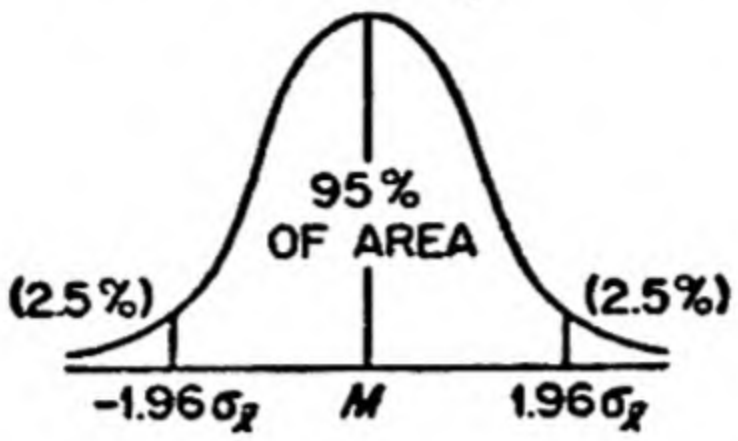


Fig. 7.4a. Distribution of means of all possible samples of same size around their universe mean.

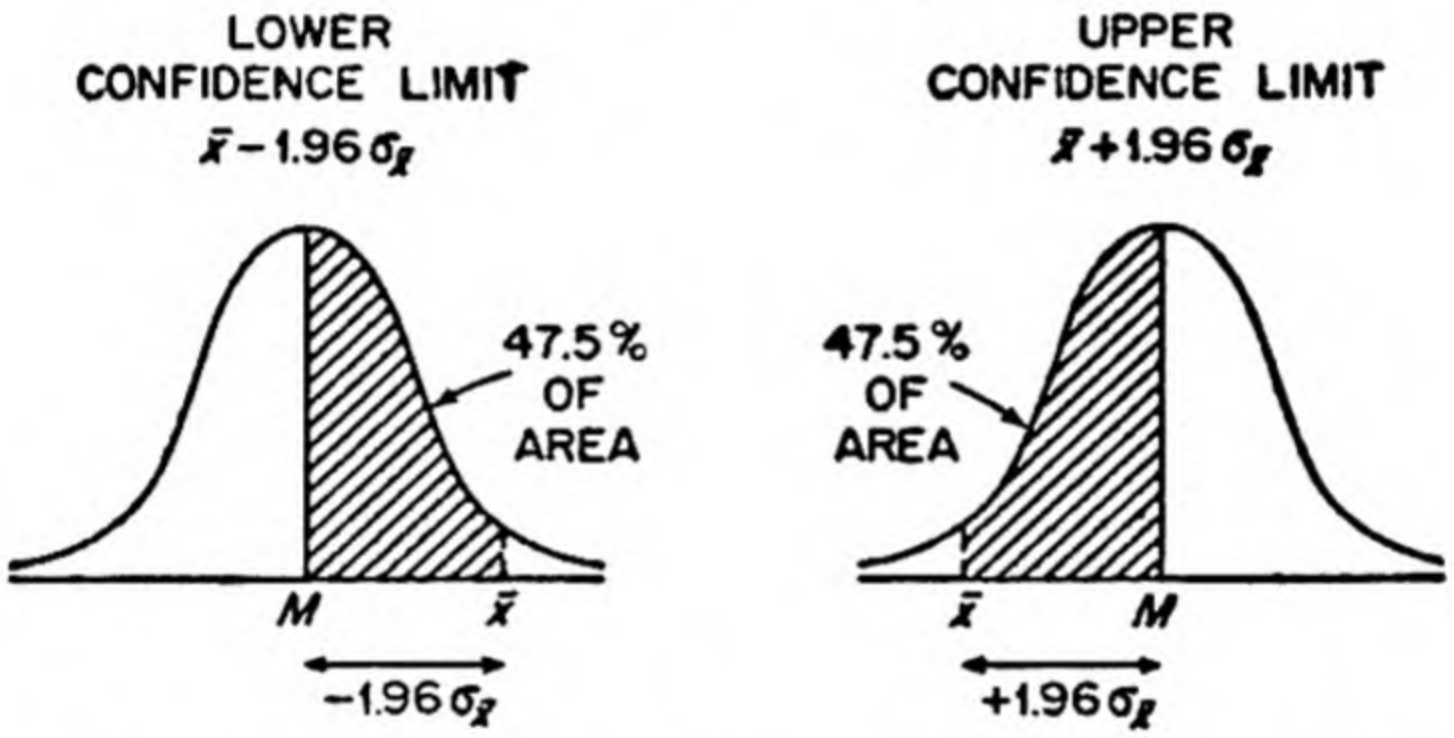


Fig. 7.4b. The mean of the sample may be above or below the universe mean.

cent confidence interval. Ninety-five per cent of the intervals obtained by this method will contain the universe mean.

$$\bar{X} \pm 1.96\sigma_{\bar{X}} = 4 \pm 1.96\ (.96) = 4 \pm 1.9$$

$$= \underline{2.1 \text{ to } 5.9} \text{ is the confidence interval}$$

The universe mean gain in IQ is estimated at some fixed value between the limits 2.1 and 5.9. (We could be wrong in 5 per cent of the samples. We are betting that our sample mean is one of the other 95 per cent.)

Note that the variability of sample means around the universe mean, as measured by the standard error, is always less than the variability of observations in a single sample around the sample mean, as measured by the standard deviation. In our problem, the stand-

ard error is ±.96; the standard deviation, ±6.7. This difference in variability is to be expected. In our single sample almost one-fourth of the 50 children changed 10 or more points in IQ over the three-year period. The distribution of *mean IQ changes* from all possible samples of 50 children would more closely approximate the universe mean change of zero than does the distribution of observations in a single sample.

Correction for Finite Sampling. The standard error formula that we have used assumes that we are selecting random samples from an indefinitely large universe. But most of our problems deal with finite universes. Our sample consisted of 50 cases out of a finite universe of 1,000. It makes little difference whether the universe is finite or indefinitely large provided that the sample is a small part of that universe. But as the sample size becomes large relative to the size of the universe, the sample values fluctuate less and less from the universe value. Consequently, the standard error formula must be corrected for finite sampling where the sample size is more than 10 per cent of the universe. The correction factor is $\sqrt{(N-n)/(N-1)}$ where N is the size of the universe and n, the size of the sample.

$$\text{est } \sigma_{\bar{X}} = \frac{s}{\sqrt{n-1}} \sqrt{\frac{N-n}{N-1}}$$

If the sample size is 100 out of a universe of 10,000, the correction factor represents a reduction in standard error of less than 1 per cent.

$$\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{9900}{9999}} \doteq \sqrt{.99} \doteq 99\%$$

But if the sample size is 5,000 out of a universe of 10,000, the correction factor represents a reduction in standard error of about 30 per cent.

$$\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{5000}{9999}} \doteq \sqrt{.50} \doteq 70\%$$

EXERCISES

1. What is the difference between the standard deviation and the standard error? What is the basis for the use of the standard error? What is a *z* score?

2. The accompanying table gives the number of school years completed by a random sample of 82 semiskilled workers in a steel plant. Could this

sample have come from a universe where the mean number of school years completed is 12?

Number of School Years Completed	*Frequency*
0– 1	3
2– 3	5
4– 5	6
6– 7	10
8– 9	15
10–11	15
12–13	16
14–15	9
16–17	3

3. In a random sample of 101 students from a certain university, the mean age is 20 years, and one standard deviation is 2.5 years. Predict the mean age of all students in this universe (with 95 per cent confidence).

4. If the mean monthly income of a random sample of 65 workers in a certain plant (hiring one thousand workers) is $300, what would you predict as the 99 per cent confidence limits within which the mean monthly income of the one thousand workers lies? The standard deviation of the sample is $100.

7.2. Testing Hypotheses about the Difference between Two Means

Instead of comparing a sample mean with a hypothetical universe mean, we may want to compare two sample means.

Problem. We know that the long-term trend in the size of the American family has been downward, and that upper-income families, best able economically to have many children, have been rearing the smallest families.

In very recent years, there seems to have been some slight increase in the number of children per family. The upper-income groups appear to lead this new trend, with a greater number of children per family than middle- or lower-income groups.

We want to know if the difference in the mean number of children per family between upper- and middle-income families is statistically significant.

Null Hypothesis. There is no difference in the mean number of children of a completed upper-income family and a completed middle-income family.

A *positive difference* between the means will indicate that the mean number of children is greater for upper-income families and a *negative*

difference will indicate that the mean number of children is greater for middle-income families.

We want to test the null hypothesis against the *alternative hypotheses* that the mean number of children per upper-income family

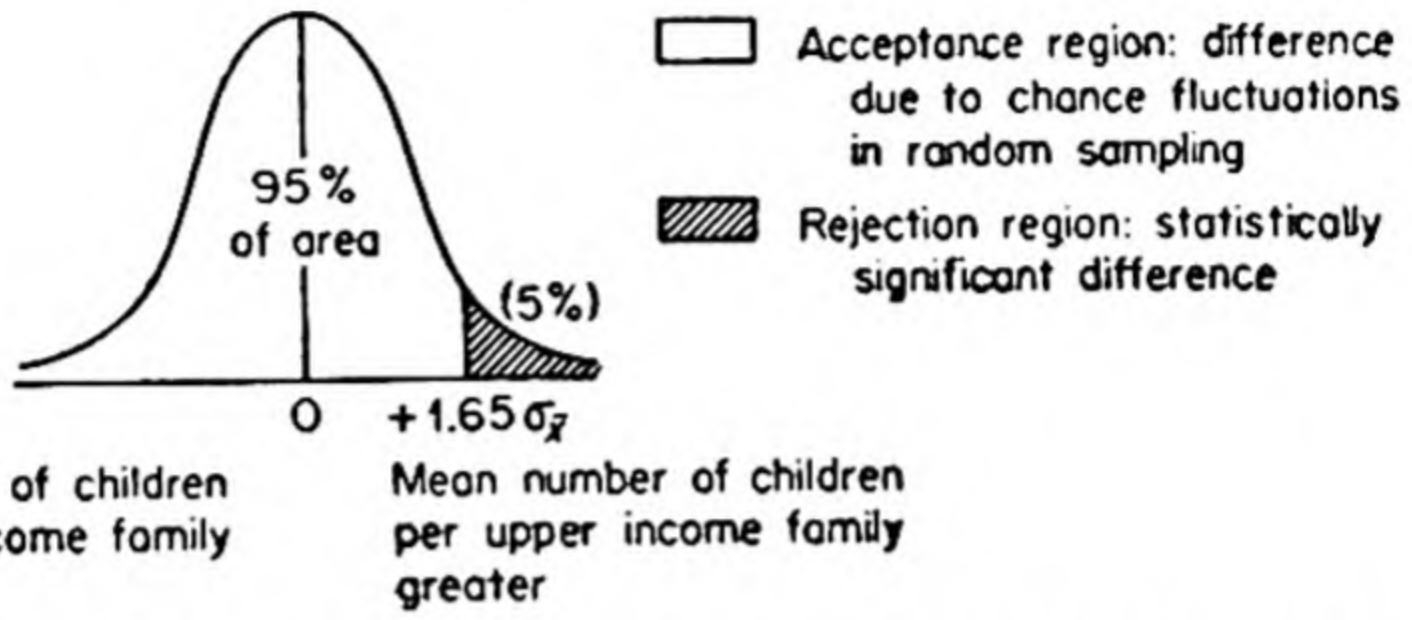


Fig. 7.5. Difference between means of all possible samples of given size from a universe with "0" difference.

is greater than the mean number per middle-income family. Hence to minimize the risk of accepting the no-difference hypothesis if it is false, we put the region of rejection of the null hypothesis in the right tail of the normal curve. Setting up a 5 per cent level of significance, we shall consider any difference between sample means less than +1.65 standard errors to have happened because of chance fluctuations in random sampling, and, consequently, not discrediting the null hypothesis. If the difference is greater than +1.65 standard errors, we will accept our alternative hypotheses that the mean number of children per upper-income family is greater than the mean number per middle-income family.

Procedure. The study is conducted in a midwestern city. We define the universe of upper-income families to include those completed families in the city whose annual income for the preceding year was $7,500 or over; the universe of middle-income families includes those completed families whose annual income was between $3,500 and $7,500. A completed family is one in which the wife is over 45. (Note that a weakness in this study lies in the fact that if this trend toward increase in number of children among upper-income families is a very recent one, it would not be reflected in those completed upper-income families where the wife is over 45.)

We select a random sample of 125 upper-income families out of a universe of 2,000, and 145 middle-income families out of a universe

of 4,000. (The data are hypothetical.) In the sample of 125 upper-income families, the mean number of children per family is 3.3, and the standard deviation is .9. In the sample of 145 middle-income families, the mean number of children per family is 2.8, and the standard deviation is 1.2. Is this difference of .5 just a chance difference due to fluctuations in random sampling? Could these two samples have come from universes having the same mean number of children per family?

Let $\bar{X}_1$ be the mean number of children per family in the sample of 125 upper-income families, and let $\bar{X}_2$ be the mean number of children per family in the 145 middle-income families. If the *means* of all possible random samples of the given size in each of the two universes are approximately normally distributed, then the *differences between the sample means* are also approximately normally distributed around the difference between the universe means. The standard error of the sampling distribution of differences between two means is

$$\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2}$$

which can be estimated at

$$\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}} \quad * \qquad (17)$$

In our problem,

$$\bar{X}_1 = 3.3, \qquad n_1 = 125, \qquad s_1 = .9$$

$$\bar{X}_2 = 2.8, \qquad n_2 = 145, \qquad s_2 = 1.2$$

$$\bar{X}_1 - \bar{X}_2 = .5$$

Is this difference significantly different from 0? We can test the

* The standard error formula above (17) is used when the two samples are independent of one another, and σ_1 may not equal σ_2. If we are testing the hypothesis that the two random samples come from the same universe, the formula for the standard error of the difference between two means is:

$$\text{est } \sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{(s_p^2)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where s_p^2, the pooled estimate of the variance, is equal to

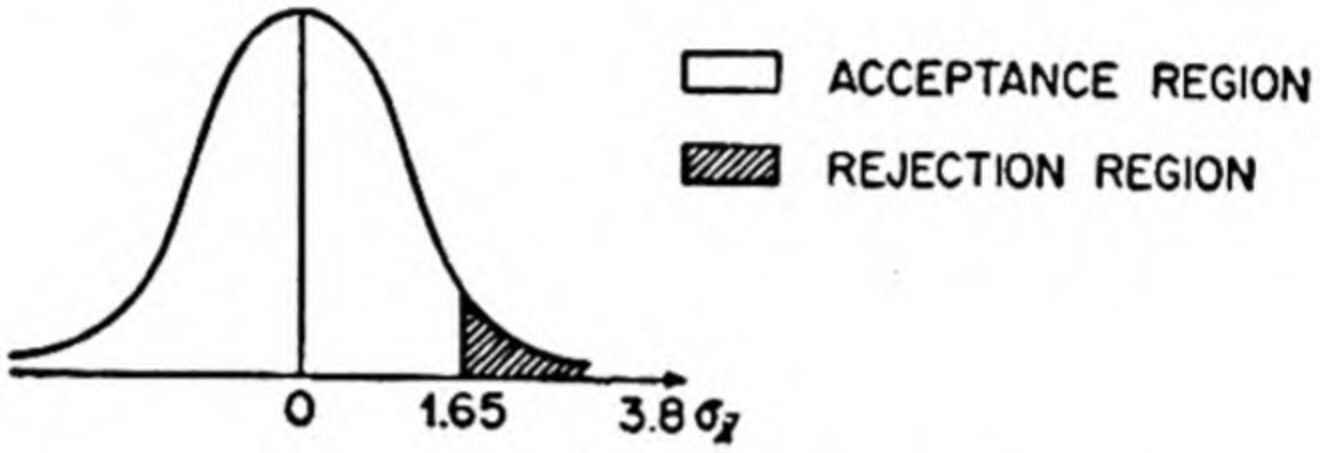
$$\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

null hypothesis of no difference by translating the raw score distance of .5 into a standard, or z score:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}}\dagger = \frac{3.3 - 2.8}{.13} = \frac{.5}{.13}$$

$$\doteq 3.8 \text{ standard error units}$$

If our null hypothesis were true that there is no difference between the mean number of children in completed upper-income families and completed middle-income families, then a positive difference as large



or larger than that found in our samples could have occurred by chance only 7 times out of 100,000. We can safely reject our null hypothesis of no difference between mean number of children in upper- and middle-income families in favor of the alternative hypotheses that the mean number of children per upper-income family is greater than the mean number per middle-income family.

7.3. The t-Test of Significance for Small Samples

We have said that the means from all possible random samples of the same size are distributed in an approximately normal fashion around the universe mean if the variable is normally distributed in its universe, or if the sample is sufficiently large. The standard error of the sampling distribution of means is $\sigma/\sqrt{n}$. It is estimated at $s/\sqrt{n-1}$ when the standard deviation of the universe is un-

† where $\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}} = \sqrt{\frac{(.9)^2}{124} + \frac{(1.2)^2}{144}} = \sqrt{\frac{.81}{124} + \frac{1.44}{144}}$

$$= \sqrt{.00653 + .01} = \sqrt{.01653}$$

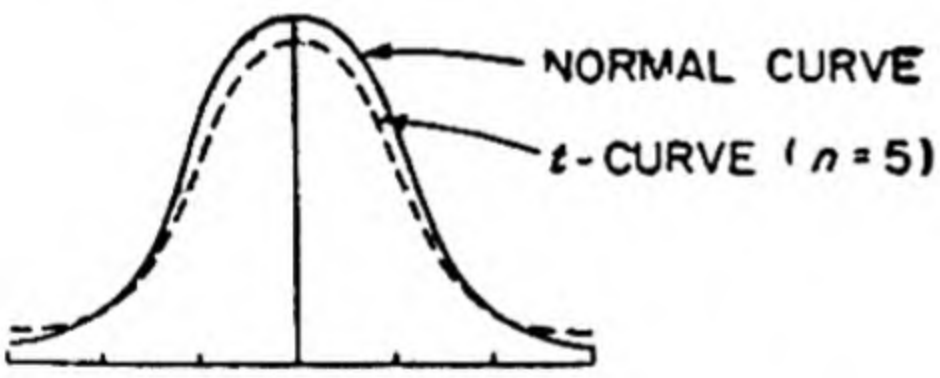
$$\doteq \underline{.13}$$

known. This is the denominator of what we have called the z score $(\bar{X} - M)/(s/\sqrt{n-1})$. The denominator of the z score will vary from sample to sample, as will the numerator. The distribution of the denominator $s/\sqrt{n-1}$ is not normal, and, consequently, the sampling distribution of the z score $(\bar{X} - M)/(s/\sqrt{n-1})$ is not normal. This departure from normality is not serious for large samples, but when samples are small (under 30) we do not assume normality for the sampling distribution of $(\bar{X} - M)/(s/\sqrt{n-1})$.

If the variable is normally distributed in the universe, or only moderately skewed, and if we calculate a z-score in which we use in our denominator the standard deviation of the *sample*, then our z-score varies according to what is called the Student-Fisher t distribution. The t distribution was discovered by W. A. Gosset, who published under the pen name of "Student." For *large* samples the t distribution and the normal distribution are the same.

$$t = \frac{\bar{X} - M}{s/\sqrt{n-1}} \quad \text{or} \quad \frac{(\bar{X} - M)}{s}\sqrt{n-1} \tag{18}$$

The sampling distribution of t depends on the number of *degrees of freedom* (which is equal to $n - 1$, or one less than the size of the sample).[1] There are different t-curves for different sample sizes. The t-curve is symmetrical and bell-shaped but has more values farther away from the mean than does the normal curve. We go out farther in standard-error units on t-curves to cover 95 per cent of the area than we do on a normal curve. When the sample size is over 30, the normal curve may be substituted for the t-curve.



Example of Use of t-Test of Significance. We want to test a hypothetical universe value of 3 as the mean number of books of fiction read per year by university students. We will set up our level of significance at 1 per cent. Our 26-student random sample mean is 2, and the standard deviation of the sample is 1.5. The t-ratio equals:

$$t = \frac{\bar{X} - M}{s}\sqrt{n-1} = \frac{2-3}{1.5}\sqrt{25} = \frac{-1}{1.5} \times 5 = \begin{matrix}-3.3 \text{ standard} \\ \text{error units}\end{matrix}$$

[1] The number of degrees of freedom in a t-test involving the difference between two means is $n_1 + n_2 - 2$.

With a sample of 26 students, $n - 1$ degrees of freedom equals 25. The *t*-table, page 169, gives the values of *t* corresponding to different degrees of freedom and to different probabilities of exceeding given values of *t*, when the region of rejection is distributed in both tails of the distribution.

We will reject the hypothesis of an average of 3 books of fiction per year read by university students if the *t*-ratio is in the region of rejection. The 1 per cent area of rejection is distributed in both tails of the distribution, .5 per cent in either tail. For 25 degrees of freedom, the table shows that there is a 1 per cent probability of exceeding a *t*-ratio of 2.787. Hence the probability of getting a *t*-ratio of 3.3 standard error units (plus or minus) would be less than 1 per cent. We can therefore reject the hypothetical universe mean of 3 books of fiction per year read by university students.

EXERCISES

1. In a sample study of 15 cities with a population of 50,000 to 500,000, the mean monthly rent was found to be $60, with a standard deviation of $10. For a random sample of 30 cities with a population between 2,500 and 5,000, the mean monthly rent was $50, with a standard deviation of $12. Is this difference in mean monthly rent between smaller and larger cities in the region of acceptance or rejection of the null hypothesis (no difference between the means)?

2. We are interested in determining the difference in participation in extracurricular campus activities between male and female members of the sophomore class of a certain university. The mean number of extracurricular campus groups to which a random sample of 25 male students belong is 2.4 (with a standard deviation of .4); the mean number for 25 female students is 2.8, with a standard deviation of .6. Test the null hypothesis (no difference between the means) at the 5 per cent and the 1 per cent level. Explain your results.

3. How large a sample would be required to say, with 95 per cent confidence, that the mean number of extra-curricular activities engaged in by sophomores lies somewhere between 2.6 and 2.8? The standard deviation of the universe has been found from a previous study to be 1.

4. An attitude test, consisting of 12 stories of actions or attitudes of non-Jews to Jews, was given to non-Jewish college and university students. The students were rated on a five-point scale according to their degree of approval or disapproval of each story, 1 indicating most unfavorable attitude, and 5, most favorable. The sum of the 12 ratings was the attitude score of each student.

(a) A random sample of 99 non-Jewish students from school *A* (where 8 per cent were Jewish students) had a mean score of 41.70, compared with the higher mean of 51.79 for a sample of 259 students from school *B*, where there were no Jews (according to the student records of religious affiliation). The $\sigma_{\bar{X}_A}$ was .481, and the $\sigma_{\bar{X}_B}$ was .571. Test the hypothesis of no difference in mean score between the two schools.

(b) The students were asked to indicate whether or not they ever had any close and intimate friends who were Jewish, and if so, how many. The mean score of the 241 students stating they had had no intimate Jewish friends was 43.61, with a $\sigma_{\bar{X}}$ of .601. The mean score of the 93 students with more than three such friends was 48.05, with a $\sigma_{\bar{X}}$ of .701. Test the hypothesis of no difference in mean score between the two groups of students. How do you explain the variation in results between (a) and (b)? (*Source:* Howard Harlan, "Factors Affecting Attitude Toward Jews," *American Sociological Review*, December 1942.)

5. To determine whether bilingualism retards mental development, a study was made of a sample of native-born Italian children, who had simultaneously learned Italian and English from infancy and through the developmental period.

The children, divided into low and high bilingual groups according to their score on a schedule designed to determine the extent of their bilingual background, were given a nonlanguage intelligence test.

Among the eleven-year-olds, can the hypothesis of no difference in intelligence between the two bilingual groups be accepted? (Compute the last four columns of the table.)

Comparison of Low and High Bilingual Groups on the Pintner Nonlanguage Test

(American-born Italian group, age eleven)

Bilingual Score	*Number*	*Mean Bilingual Score*	*Mean Intelligence Score*	*Standard Deviation*
Low (0–6)	45	2.87	292.60	65.20
High (18–35)	55	23.78	282.16	76.60

Bilingual Score (*cont'd*)	$\sigma_{\bar{X}}$	$\sigma_{\bar{X}_1-\bar{X}_2}$	$\bar{X}_1 - \bar{X}_2$	$z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}}$
Low (0–6)				
High (18–35)				

Source: Seth Arsenian, *Bilingualism and Mental Development* (Bureau of Publications, Teachers College, Columbia University, 1937), p. 106.

7.4. A Proportion: Testing Hypotheses and Making Estimations

Problem. The time is one month before a presidential election. We are interested in knowing whether the University of Michigan student body favors the Republican or Democratic candidate. It is impractical to poll the student body as a whole; we shall poll only one random sample from this universe.

Hypothesis. In the Michigan student body, one-half of those who favor either of the two major political candidates are for the Republican candidate.

We set our significance level at 5 per cent. The alternative to the hypothetical proportion of .50 may be either greater or less than .50. Hence in distinguishing chance fluctuations from significant difference we shall use both tails of the normal distribution.

Procedure. Using a table of random numbers, we select a sample of 115 students from a card-catalogue listing of the Michigan student body; each card is numbered. Each person selected is asked his candidate preference. Our sample results are:

Candidate Preference	*Frequency*	
Republican	61	($n = 100$ students; this is the only part
Democrat	39	(of sample we are interested in.)
Others	15	
Total:	115	

In this sample, 61 per cent of those students whose political preferences are for either of the two major candidates are for the Republican candidate. Does this refute our hypothesis that 50 per cent of the "committed" Michigan student body are for the Republican candidate? Could our sample proportion of .61 have come from a universe where the proportion is .50? What is the probability of getting a random sample proportion as far away or farther away than .61 if the actual universe proportion is .50?

The binomial distribution gives the chance of getting each of the various possible proportions for the Republican candidate. All of the various possible proportions make up the sampling distribution of proportions. The possible proportions can be denoted by X/n,

where X includes all the integers from 0 to n, and n is the size of the sample. These proportions range from 0 to 1. In a sample of 100 students, the possible proportions for the Republican candidate are $\frac{0}{100} = 0, \frac{1}{100}, \frac{2}{100}, \frac{3}{100}, \cdots \frac{98}{100}, \frac{99}{100}, \frac{100}{100} = 1$. The mean of the sampling distribution of proportions is p, and the standard deviation of the sampling distribution of proportions is $\sqrt{\frac{pq}{n}}$.

Mean of the Sampling
Distribution of Proportions $= p = .50$ (hypothetical universe probability of success)

Standard Deviation of the
Sampling Distribution of
Proportions (i.e., standard
error of a proportion) $= \sigma_{p_s} = \sqrt{\frac{p_u q_u}{n}}$ (19)

where p_u = hypothetical universe probability for the Republican candidate

$q_u = 1 - p$ = hypothetical universe probability for the Democratic candidate

n = size of sample

(Note that p_u and q_u are universe proportions and not sample proportions.)

$$\sqrt{\frac{p_u q_u}{n}} = \sqrt{\frac{(.50)(.50)}{100}} = \sqrt{\frac{.25}{100}} = \frac{.5}{10} = .05$$

The formula $\sqrt{p_u q_u / n}$ is called the *standard error of a proportion*. It is the standard deviation of the sampling distribution of all possible proportions in a binomial distribution around the mean proportion. When the sample is fairly large and p is not too small (i.e., np is greater than 5), we can approximate the binomial distribution by means of a normal curve which has a mean equal to p and a standard deviation equal to $\sqrt{pq/n}$. We shall use the normal curve as an approximation to the binomial distribution in our example.

Let us examine what we are doing: we are testing the hypothetical universe proportion of .50 for the Republican candidate against the alternative proportion different from .50 by determining how likely we are to get a sample proportion of .61 in such a universe. If it is highly unlikely, we reject the null hypothesis that .50 are for the

Republican candidate. If it is highly likely that we could get a sample proportion of .61 in a universe where the proportion for the Republican candidate is .50, the difference between .61 and .50 probably occurring because of the chance fluctuations of random sampling, then we would not reject the 50-per-cent-for-the-Republican-candidate hypothesis. The probability can be determined by finding out how far .61 is from .50 under the normal curve, in standard error units.

$$z = \frac{p_s - p_u}{\sigma_{p_s}} = \frac{.61 - .50}{.05} = \frac{.11}{.05} = 2.2 \text{ standard errors}$$

(p_s = sample proportion for Republican candidate)

If the universe proportion for the Republican candidate is .50, then .61, the sample proportion, is 2.2 standard errors away from the universe proportion. We could have a sample proportion more than 2.2 standard errors away from the universe proportion in only 2.78 per cent of the samples (2 × 1.39 per cent), or about one time out of 35. (See Normal Curve Table, page 168.)

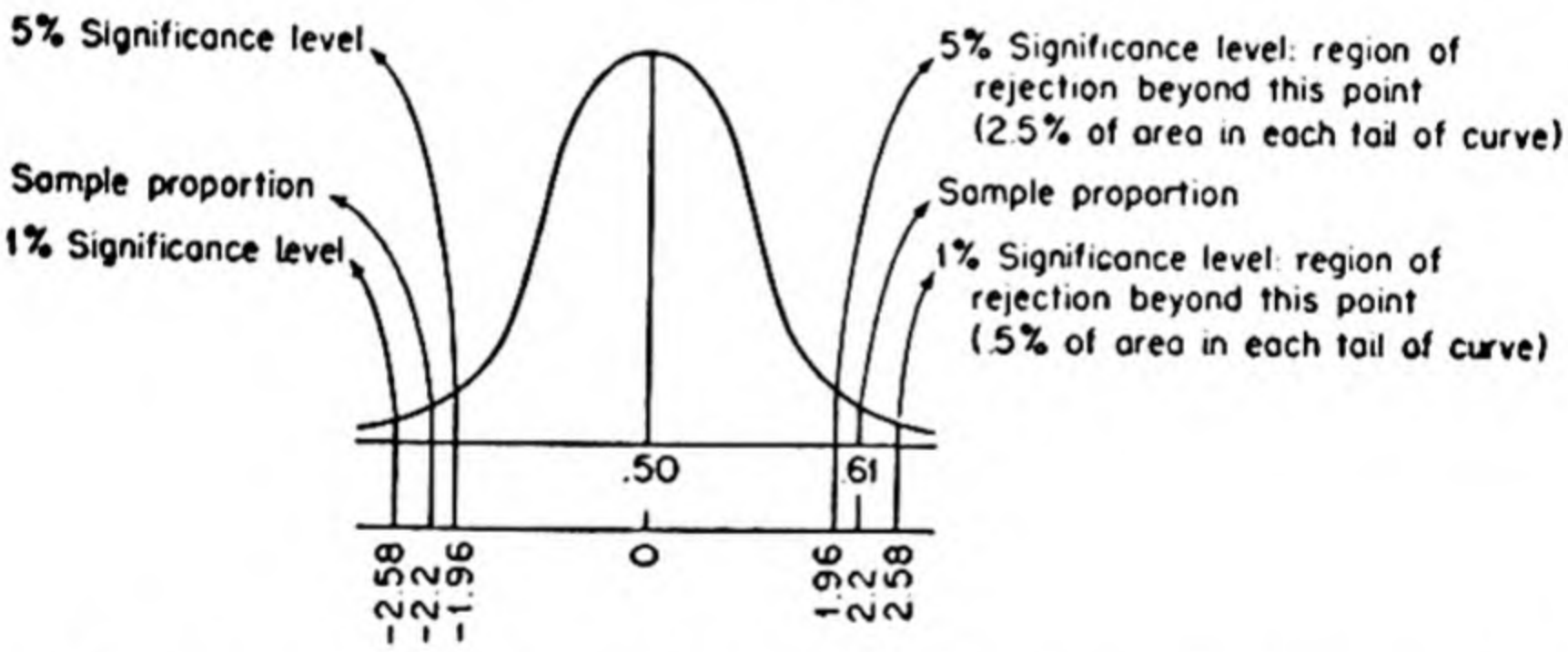


Fig. 7.6. Normal curve distribution of sample proportions around hypothetical universe proportion.

At the 5 per cent level of significance (corresponding to a *z*-score of 1.96 standard errors), we reject the hypothesis of .50 for the Republican candidate. Our sample gives a *z*-score of 2.2 standard errors away from the hypothetical universe proportion of .50, and can have happened by chance in only 2.78 per cent of the random samples from this universe.

If, before the study, we had set our level of significance at 1 per cent (with a corresponding *z*-score of 2.58 standard errors), we would not reject the hypothesis that 50 per cent of the students in the Michigan

student body are for the Republican candidate, since our difference of .11 between sample and hypothetical universe proportion may have happened by chance.

Making Estimates from a Sample Proportion to a Universe Proportion. Instead of using our sample as the basis for a statistical test between a null hypothesis and alternative hypotheses, we can use our sample proportion to estimate a confidence interval containing the universe proportion.

In our sample, .61 are for the Republican candidate. We want to estimate the confidence interval that has a 95 per cent chance of containing the universe proportion for the Republicans. We know that 95 per cent of the time the z-score distance between sample proportion and universe proportion will lie between -1.96 and $+1.96$ standard errors. Consequently, we could solve mathematically for the limits of the confidence interval containing the universe proportion. The 95 per cent confidence limits of p_u have been worked out in Table IV of the Appendix for samples of size 10, 15, 20, 30, 50, 100, 250, and 1,000. The sample proportion is given on the horizontal scale. We extend a vertical line from the sample proportion (p_s) of .61 through the two points on the belt-shaped curve that apply to our sample size of 100, one representing the lower, the other, the upper confidence limit. We then read off the corresponding values on the vertical p_u scale. Our 95 per cent confidence interval extends from about .51 to .70.

Note from Table IV that for any given size sample, the confidence interval is largest when p_s approximates .50. Also note that as the sample size increases, the confidence interval gets smaller.

If our sample is large, and neither p nor q is small, we can substitute p_s for p_u in the standard error formula in order to estimate a 95 per cent confidence interval.

$$\text{est } \sigma_{p_s} = \sqrt{\frac{p_s q_s}{n}} = \sqrt{\frac{(.61)(.39)}{100}} = \sqrt{\frac{.2379}{100}} = \sqrt{.002379} = .049$$

We can then set up an estimated 95 per cent confidence interval. We say that we are 95 per cent sure that the universe proportion is at some fixed point within the confidence interval $.61 \pm 1.96\sigma_p$, or that the confidence interval, .514 to .706, includes the universe proportion.

$$95\% \text{ confidence interval} = .61 \pm 1.96\ (.049)$$
$$= .61 \pm .096 \quad \text{or} \quad .514 \text{ to } .706$$

The substitution of the sample proportion in the standard error formula for the universe proportion (either actual or hypothetical) means that each random sample of a given universe will have a different standard error. The standard error formula is insensitive to small changes in p and q. For example, when the sample size is 100 and p equals .50, the standard error is .05; if p equals .40, the standard error is .049. But with p equal to .10 (and the sample size still 100), the standard error is .03. The σ_p is largest when p equals .50.

Change in Standard Error with Change in Value of p
(Sample size = 100)

If $p = .50$, $\sigma_p = .05$
If $p = .40$, $\sigma_p = .049$
If $p = .10$, $\sigma_p = .03$

$$\sigma_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(.5)(.5)}{100}} = \sqrt{\frac{.25}{100}} = \frac{.5}{10} = \frac{1}{20} = \underline{.05}$$

$$\sigma_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(.4)(.6)}{100}} = \sqrt{\frac{.24}{100}} = \frac{.49}{10} = \underline{.049}$$

$$\sigma_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(.1)(.9)}{100}} = \sqrt{\frac{.09}{100}} = \frac{.30}{10} = \underline{.03}$$

Effect of Sample Size on Size of Confidence Interval. As the size of the sample increases, the interval estimated to contain the universe proportion gets smaller. Assume that different size samples show .50 for the Republican candidate. On the basis of each sample we want to estimate an interval which has 95 per cent chance of including the universe proportion. With a sample of 20 we can estimate a confidence interval with a range of 46 points. With a sample of 100, we can estimate the range at 20 points. And with a sample of 1,000, the range will be 6 points.

Random Sample Size	*Approximate 95% Confidence Interval* ($p_s = .50$)
20	.27 to .73
100	.40 to .60
1,000	.47 to .53

The range of the confidence interval is not determined by the size of the universe, unless the sample represents a substantial proportion of the universe. Nowhere in the standard error formula, except in

student body are for the Republican candidate, since our difference of .11 between sample and hypothetical universe proportion may have happened by chance.

Making Estimates from a Sample Proportion to a Universe Proportion. Instead of using our sample as the basis for a statistical test between a null hypothesis and alternative hypotheses, we can use our sample proportion to estimate a confidence interval containing the universe proportion.

In our sample, .61 are for the Republican candidate. We want to estimate the confidence interval that has a 95 per cent chance of containing the universe proportion for the Republicans. We know that 95 per cent of the time the z-score distance between sample proportion and universe proportion will lie between -1.96 and $+1.96$ standard errors. Consequently, we could solve mathematically for the limits of the confidence interval containing the universe proportion. The 95 per cent confidence limits of p_u have been worked out in Table IV of the Appendix for samples of size 10, 15, 20, 30, 50, 100, 250, and 1,000. The sample proportion is given on the horizontal scale. We extend a vertical line from the sample proportion (p_s) of .61 through the two points on the belt-shaped curve that apply to our sample size of 100, one representing the lower, the other, the upper confidence limit. We then read off the corresponding values on the vertical p_u scale. Our 95 per cent confidence interval extends from about .51 to .70.

Note from Table IV that for any given size sample, the confidence interval is largest when p_s approximates .50. Also note that as the sample size increases, the confidence interval gets smaller.

If our sample is large, and neither p nor q is small, we can substitute p_s for p_u in the standard error formula in order to estimate a 95 per cent confidence interval.

$$\text{est } \sigma_{p_s} = \sqrt{\frac{p_s q_s}{n}} = \sqrt{\frac{(.61)(.39)}{100}} = \sqrt{\frac{.2379}{100}} = \sqrt{.002379} = .049$$

We can then set up an estimated 95 per cent confidence interval. We say that we are 95 per cent sure that the universe proportion is at some fixed point within the confidence interval $.61 \pm 1.96\sigma_p$, or that the confidence interval, .514 to .706, includes the universe proportion.

$$95\% \text{ confidence interval} = .61 \pm 1.96\ (.049)$$
$$= .61 \pm .096 \quad \text{or} \quad .514 \text{ to } .706$$

We can now say that we are *95 per cent* confident of predicting an election *within* 1 per cent on either side of the estimated proportion with a random sample of 9,604.

If we think that p_u is likely to approximate .80, we would need a sample of 6,147 to be 95 per cent sure of predicting an election within 1 per cent on either side of the estimated proportion.

$$\pm 1.96\sigma_p = \pm .01$$

$$(1.96)\sqrt{\frac{(.80)(.20)}{n}} = .01$$

$$(1.96)\sqrt{\frac{.16}{n}} = .01$$

$$(1.96)\left(\frac{.40}{\sqrt{n}}\right) = .01$$

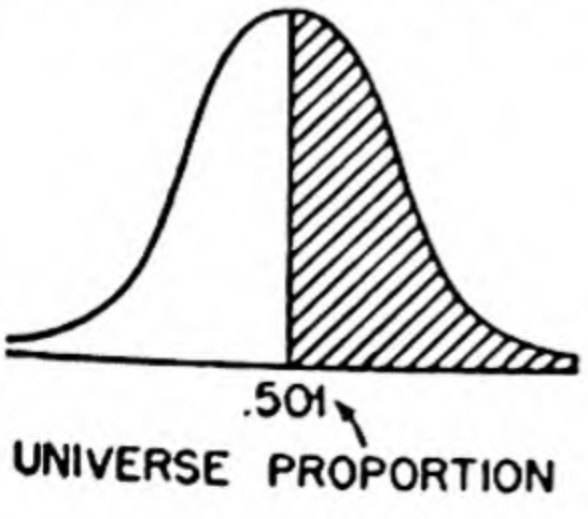
$$\frac{.784}{\sqrt{n}} = .01$$

$$.784 = .01\sqrt{n}$$

$$78.4 = \sqrt{n}$$

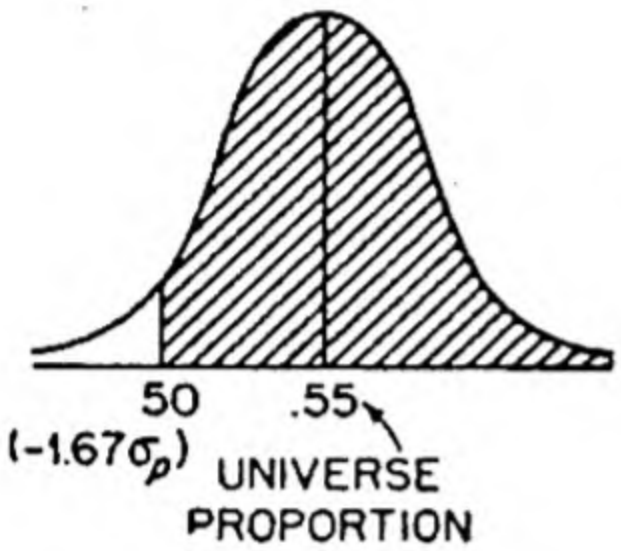
$$6{,}147 \doteq n$$

How Small Should Standard Error Be? How large a standard error will we be satisfied with? If we are trying to predict an election, the difference between 49 per cent and 51 per cent may be crucial. Both errors of chance and errors of bias can occur in predicting election returns. Under the unlikely condition that there are no errors of bias at all, if one standard error equals 3 per cent and the Republican proportion of the two-party vote in a certain state is 50.1, then a pollster's chance of correctly predicting the outcome in the state is about 50-50. With this universe proportion of just about 50 per cent, one-half the possible random sample proportions will be above, and one-half below 50 per cent. And this 50-50 chance of correctly predicting the results assumes no errors of bias at all.



Fifty per cent chance of getting a sample proportion in the part of the curve above .50, and thus correctly predicting the election

When the Republican vote in a state is 55 per cent, then a poll with a standard error of 3 per cent has a 95 per cent chance of correctly predicting the winner. With the universe proportion at .55, 95 per cent of the possible random samples will give a Republican proportion above .50.



Ninety-five per cent chance of getting a sample proportion above .50 from this universe, and thus correctly predicting the outcome

$$\frac{z = P_s - P_u}{\sigma_{p_s}} = \frac{.50 - .55}{.03} = -1.67 \text{ standard errors}$$

EXERCISES

1. The election returns for a certain city gave .40 of the votes to the Democratic nominee and .60 to the Republican nominee. What is the standard error of the proportion for random samples of 25 drawn from this city? Samples of 100? 400? 1,600? What effect does changing the size of the sample have on the standard error?

2. If 30 per cent of all respondents are in occupations similar to their fathers, test the hypothesis of no difference in this characteristic between all respondents and professional men, 25 per cent out of a sample of 150 professional men being in occupations similar to their fathers?

3. A sample study of a certain city indicated that .50 of the families were home owners. The standard error is .05.

(a) Estimate the per cent of home owners for the city. (Set up a 95 per cent confidence interval.)

(b) What is the size of the sample? (Solve for n in the standard error formula.)

(c) On the basis of the sample, can we say that a greater proportion of families are home owners in this city than in a nearby city, where .45 are home owners?

7.5. Testing Hypotheses for the Difference between Two Sample Proportions

Problem. We want to know whether one of the factors perpetuating class differences in our society, or, at least, accompanying class difference is a different level of aspiration among adolescents of various classes.

Null Hypothesis. There is no difference in the proportion of adolescents with professional and business career aims in lower-class families from the proportion in middle-class families. We want to test this null hypothesis against the alternative hypotheses that the proportion with such aims in middle-class families is greater than that in lower-class families.

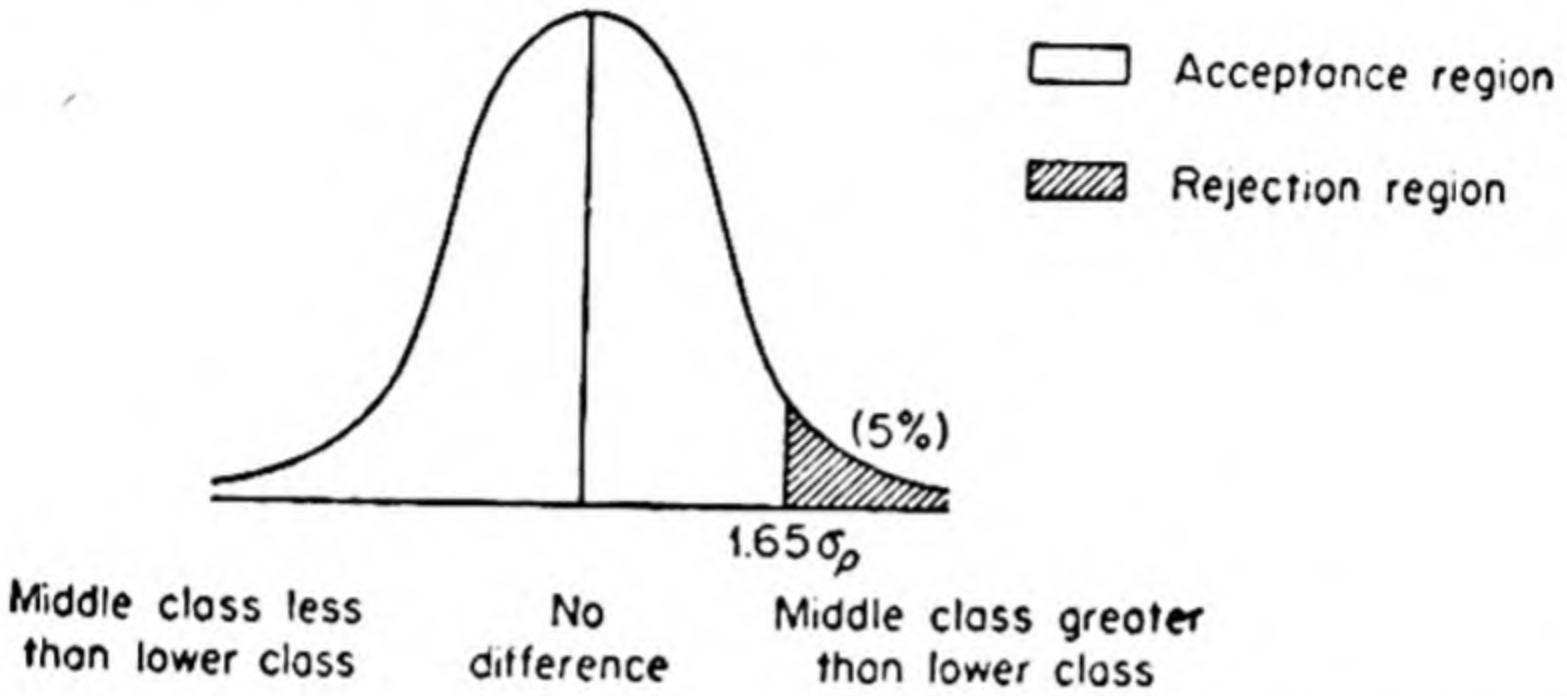


Fig. 7.7. Difference between proportions in all possible samples of given size from universe with no difference.

We set up a 5 per cent significance level and try to minimize the risk of accepting a false null hypothesis by putting the region of rejection in the right tail of the normal curve. If the proportion of middle-class adolescents with professional aims exceeds the lower class by a difference so great that it could have occurred in only 5 samples out of 100 in a universe where there is no difference between classes, then we reject the null hypothesis in favor of the alternative hypotheses.

Procedure. The study is conducted in a small midwestern town. It is found that 67 out of a random sample of 200 adolescents of the middle class have business or professional aims, and 25 out of a sample of 250 adolescents of the lower class have such goals.

The population of adolescents from the ninth to twelfth grade in high school is sampled. Middle- and lower-class families are defined according to the father's occupation and family income.

The proportion of middle-class adolescents having business or professional aims is .33, and the proportion of lower class adolescents having such aims is .10. The difference between the two proportions is .23. Is this difference sufficiently great to warrant our accepting the alternative hypotheses, or is it in the range within which we will accept the null hypothesis? How often could we have gotten samples

with this difference if there is actually no difference in professional aims between middle- and lower-class adolescents?

Let p_{s_1} be the proportion of adolescents in the middle-class sample who have business and professional aims, and p_{s_2} be the proportion of adolescents in the lower-class sample with such aims:

$$p_{s_1} - p_{s_2} = .33 - .10 = .23$$

Is a difference of .23 in the acceptance region of the null hypothesis? To determine the probability of getting a difference this large from random samples of our hypothetical universe, we convert the difference into z-score form. And to convert to z-score form, we must know the standard error for the difference between two proportions. If n_1 and n_2 are sufficiently large, the difference between the two sample proportions will be approximately normally distributed with a mean difference equal to the difference between the universe means and a standard error equal to

$$\sigma_{p_{s_1} - p_{s_2}} = \sqrt{\sigma_{p_1} + \sigma_{p_2}} = \sqrt{\frac{p_{u_1}q_{u_1}}{n_1} + \frac{p_{u_2}q_{u_2}}{n_2}} \tag{20}$$

Since our null hypothesis states that there is no difference in proportion with professional aims between the two classes ($p_{u_1} = p_{u_2}$), the standard error can be stated as:

$$\sigma_{p_{s_1} - p_{s_2}} = \sqrt{p_u q_u \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{21}$$

where p_u, the proportion with professional aims in the hypothetical universe, is estimated by combining the two samples of adolescents into one sample of 450, in which 92 have business and professional aims. We estimate p_u at $\frac{92}{450}$, or about 20 per cent. The standard error of the difference between two proportions is .04:

$$\sigma_{p_{s_1} - p_{s_2}} = \sqrt{p_u q_u \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{(.2)(.8)\left(\frac{1}{200} + \frac{1}{250}\right)} = .04$$

The z-ratio is equal to 5.8 standard errors:

$$z = \frac{p_{s_1} - p_{s_2}}{\sigma_{p_{s_1} - p_{s_2}}} = \frac{.23}{.04} \doteq 5.8 \text{ standard errors}$$

A difference this large or larger could have happened by chance an infinitesimally small number of times. We therefore reject the null hypothesis that there is no difference in professional aims between

middle- and lower-class adolescents in favor of the alternative hypotheses that a greater proportion of middle-class adolescents have such aims. Our statistical computation does not tell us why there is a difference—whether it is a difference in anticipated opportunity, perhaps, rather than a difference in desire. Nor do our study results from a small midwestern town necessarily apply to urban areas.

REFERENCES

Dixon, W. J., and F. J. Massey, *Introduction to Statistical Analysis*, chaps. 7 and 9. New York: McGraw-Hill Book Company, Inc., 1951.

Freund, J. E., *Modern Elementary Statistics*, chaps. 9 and 10. New York: Prentice-Hall, Inc., 1952.

McNemar, Quinn, *Psychological Statistics*. chap. 5. New York: John Wiley & Sons, Inc., 1949.

Mode, Elmer B., *Elements of Statistics*, Rev. ed. chaps. 9 and 12. New York: Prentice-Hall, Inc., 1951.

EXERCISES

1. In the last congressional primaries, a random sample of 200 registered voters in Maine and 250 in New York were asked the question: With what party are you affiliated? In Maine, sixty per cent of the sample said they had Republican affiliations; fifty-five per cent in New York declared similar affiliations. Would you say that the difference in the proportion of Maine and New York Republican affiliators is sufficiently great to warrant our accepting alternative hypotheses, or is it in the range within which we will accept the null hypothesis?

2. In a study of parental rejection it was found that in 100 families where one or both of the spouses distrusted each other, 85 of the families had rejected their children. Among 100 families where there was no parental distrust, 30 had rejected their children. Is this difference in rejection of children between families where there is distrust of spouse and families with no such distrust sufficiently great to warrant our accepting alternative hypotheses, or can we accept the null hypothesis?

3. In answer to the question: are you satisfied with your present living accommodations?, 85 per cent of a sample of 450 owners answered in the affirmative, as did 50 per cent of the 125 sampled tenants. Test the null hypothesis. (*Source:* Theodore Caplow, "Home Ownership and Location Preferences in a Minneapolis Sample," *American Sociological Review*, December 1948.)

4. If 10 per cent of a sample of 150 neurotics engage in excessive smoking, and 5 per cent of a sample of 200 normal persons smoke excessively, does this difference in excessive smoking between the two groups invalidate the null hypothesis?

5. Assume that the data in the accompanying table come from a random sample of higher and lower socio-economic groups in Steubenville, Ohio. At the higher socio-economic level, test the hypothesis of no difference in the per cent of home owners between respondents with native-born fathers and respondents with foreign-born fathers. Do the same at the lower socio-economic level. Can you give a sociological interpretation to the results?

Home Ownership among Persons of Foreign Extraction in Steubenville, Ohio

	Higher Socio-Economic Group		Lower Socio-Economic Group	
	Percentage of Home Owners	*Total Cases*	*Percentage of Home Owners*	*Total Cases*
Respondents whose fathers were born:				
in the United States	54%	(96)	35%	(85)
outside the United States	74%	(109)	59%	(71)

Source: John P. Dean, "The Ghosts of Home Ownership," *The Journal of Social Issues*, VII Nos. 1 and 2 (1951), 62.

7.6. The Chi-Square Test of Significance

In the following section, a null hypothesis stating that there is no difference in the proportion losing authority between employed and unemployed fathers is tested with the familiar z-score. Then, after a brief introduction to the use of contingency tables, the same hypothesis is retested with chi-square.

Problem. Within the last twenty-five years there have been periods of mass unemployment in the United States. What happens to the authority of the father within the family when he loses his job? We shall want to compare the change in family authority among unemployed fathers with the change among employed fathers over the same period of time, so that we can be sure there is nothing in the culture making for a general decline in parental authority. Over a 12 to 18 month period, do unemployed fathers lose authority in their family more frequently than employed fathers?

Null Hypothesis. We set up our null hypothesis that there is no difference in the proportion losing authority between employed and unemployed fathers. Over a 12 to 18 month period, employed fathers are just as likely to lose authority as unemployed fathers.[2]

We shall set up a 5 per cent level of significance.

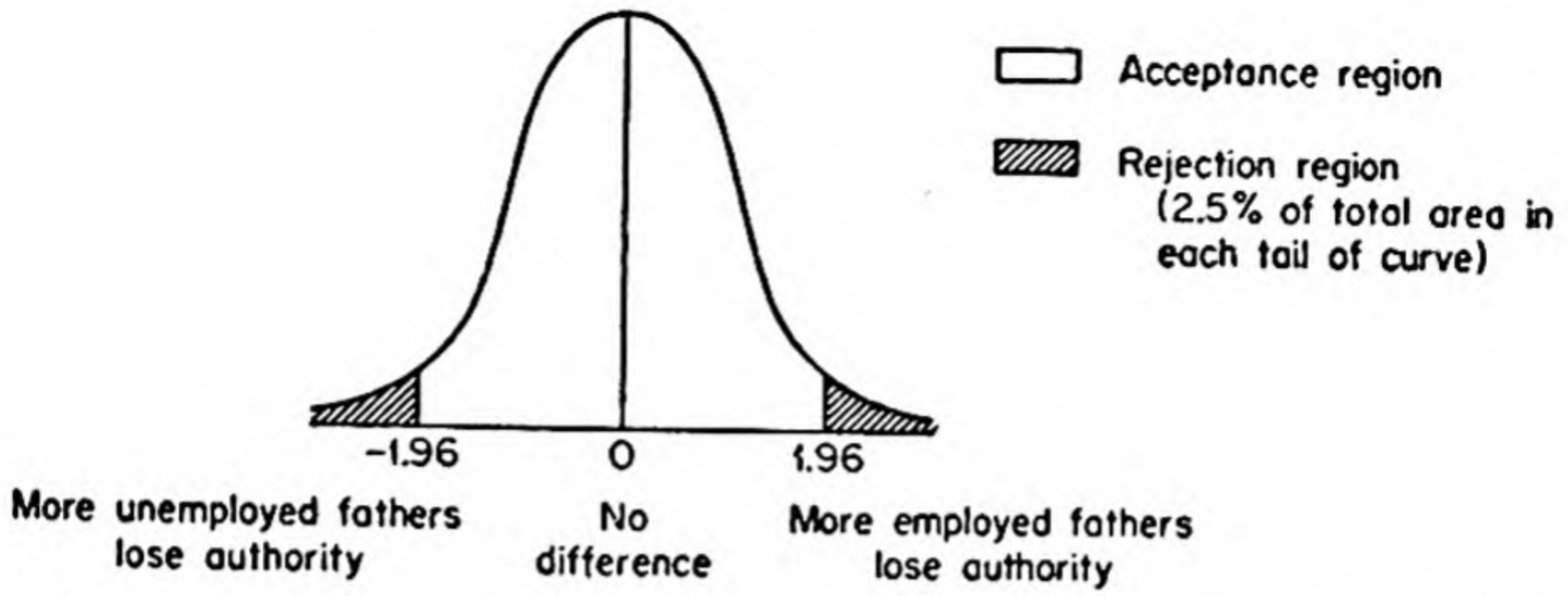


Fig. 7.8. Distribution between proportions among all possible samples of given size in universe with no difference.

Procedure. Assume that from 1 to 1½ years before the study a recession hit the Detroit automobile industry, which resulted in large-scale unemployment.

A random sample of 1,200 Detroit fathers is selected. We find out whether the men are employed or unemployed. We also try to determine by psychological test and depth interview with each member of the 1,200 families what happened to the authority of the father in his family within the past 12 to 18 months.

There are four possibilities with regard to authority status. Fathers may have (1) gained authority; (2) lost authority; (3) retained authority; or (4) had no authority at the beginning or the end of the period. We are interested in groups (2) and (3)—those who lost and those who retained authority. We consequently include in our final sample only those who retained or lost authority.

We include only those who are employed or, if unemployed, lost their job from 12 to 18 months ago. Thus we want unemployment to precede change in family authority. We want the cause-effect relationship, if any, to go from employment status to change in authority.

In our random sample of 1,200 fathers let us assume that 1,000 can fit into the desired categories: they are either employed or lost their job 12 to 18 months ago; and they either retained or lost authority in their family over the same period. Among the 1,000, 800 are employed and 200 are unemployed.

In the sample of 800 employed fathers, 320 (or 40 per cent) lost

[3] This hypothesis, treated differently, is taken from Stouffer and Lazarsfeld, *Research Memorandum on the Family in the Depression*, Bulletin 29 (Social Science Research Council, 1937) Appendix A.

authority in their family in the past 12 to 18 months. Among the 200 unemployed fathers, 100 (or 50 per cent) lost authority. Can we reject the hypothesis of no difference in the loss of authority between employed and unemployed fathers (at the 5 per cent level of significance)? The standard error of the difference between the two proportions is .039

$$\sigma_{p_{s_1}-p_{s_2}} = \sqrt{p_u q_u \left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = \sqrt{(.42)(.58)\left(\frac{1}{800}+\frac{1}{200}\right)} = .039$$

where the estimated universe proportion of fathers who lost authority is .42:

$$\text{est } p_u = \frac{320+100}{1000} = .42$$

and the estimated universe proportion of fathers who retained authority is .58:

$$\text{est } q_u = \frac{480+100}{1000} = .58$$

We set up the z score to test the significance of the difference between the two proportions (.40 and .50).

$$z = \frac{p_{s_1}-p_{s_2}}{\sigma_{p_{s_1}-p_{s_2}}} = \frac{.40-.50}{.039} = \frac{-.10}{.039} = -2.56 \text{ standard errors}$$

The distance between the two proportions is −2.56 standard errors. An absolute (plus or minus) difference this large could have occurred

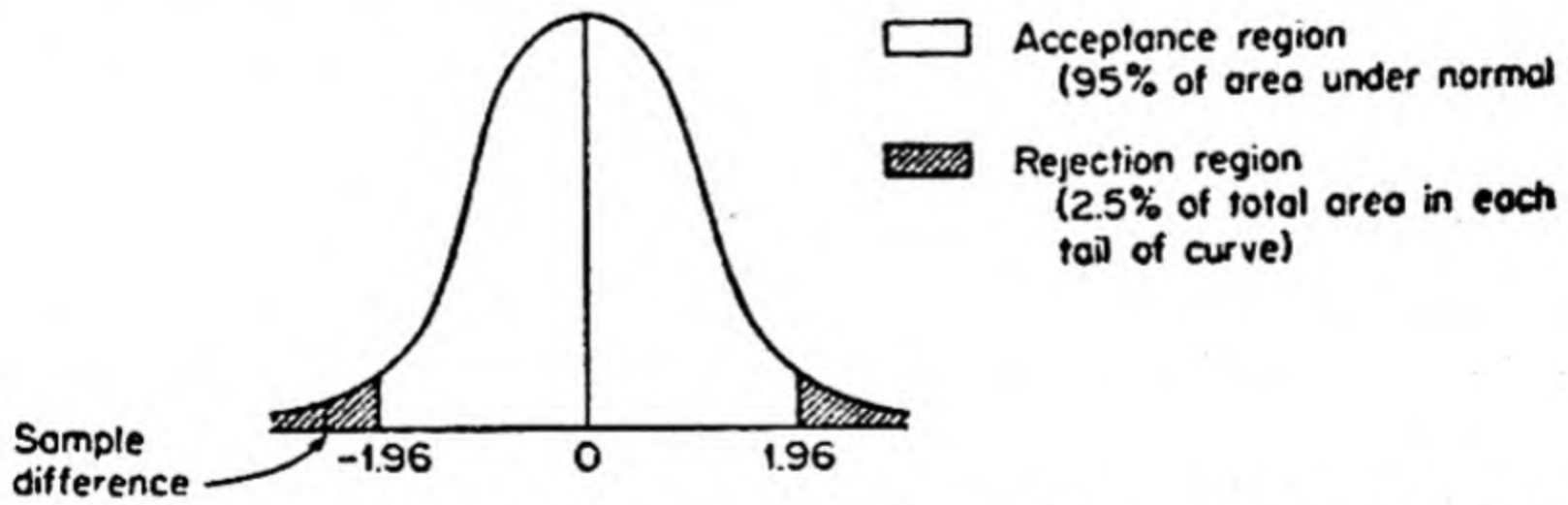


Fig. 7.9. Difference between proportions from all possible samples of given size in universe with no difference.

in only 1.04 per cent of the samples from a universe with no difference between employed and unemployed fathers who lost authority; it is therefore significant at the 5 per cent level. We reject the hypothesis that there is no difference in the loss of authority between employed and unemployed fathers.

There is another procedure for arriving at the same results. Our

study contains two attributes: (1) father's employment status; and (2) father's change in authority. We want to know whether there is *any association between* these two *attributes*. We can classify the attributes: (1) "father employed" and "father unemployed"; (2) "father retains authority" and "father loses authority."

The sample data on these two attributes might be illustrated through the use of a contingency table. Contingency refers here to the association between attributes. Since there are two dichotomous classifications in the problem, we shall need a 2 × 2 (2 rows and 2 columns) or four-fold contingency table, having four cells in addition to margin totals. The sample data can be placed in the proper cells.

Employment Status and Change in Authority: Observed Distribution of 1,000 Fathers

(Hypothetical data)

	(Column 1) *Father Unemployed*	(Column 2) *Father Employed*	
(Row 1) Father retains authority	100	480	580
(Row 2) Father loses authority	100	320	420
	200	800	1,000

This 4-fold contingency table gives the *observed distribution*. (The 4 stands for the number of cells.)

Note that *numbers* and not percentages or proportions are placed in the contingency table. The contingency table shows that the proportion of *employed* fathers losing authority is 320 ÷ 800, or .40; the proportion of unemployed fathers losing authority is 100 ÷ 200, or .50. Since these two proportions differ rather substantially, we would be inclined to think that the two characteristics, employment status and authority, are not independent of each other, but are associated. We shall not trust to the observations of one sample, however, but shall test statistically to determine whether employment and authority are independent in the universe from which the sample was drawn.

Two attributes are considered *independent* if the probability of the occurrence of one attribute is the same whether or not a second attribute occurs. The probability of a father losing authority is the same whether he is employed or unemployed. If the two attributes are independent, the proportions in each row and in each column of the contingency table are the same. That is, if the proportion of

employed fathers losing authority were .40, the proportion of unemployed fathers losing authority would also be .40. Or, if the employed fathers represented .75 of those losing authority, they would also represent .75 of those retaining authority. It is not the numbers, but the proportions that have to be the same for two attributes to be considered independent of one another.

The null hypothesis says that there is no association between a father's employment status and his authority in the family. The proportion of fathers who lose authority in the sample of 1,000 is .42. If the father's employment status and authority are independent, .42 of the 800 employed fathers and .42 of the 200 unemployed fathers would lose authority. With the margin totals of 800 and 200 remaining fixed, 42 per cent of 800 (336) and 42 per cent of 200 (84) give the number of employed and unemployed fathers who lose authority under the assumption of independence between employment status and authority in the family.

We use the marginal totals to find the expected frequency in any cell under the assumption of independence between characteristics. Out of 1,000 fathers, 420, or 42 per cent of the total sample fathers, lose authority. If the two attributes are independent, 42 per cent of the 800 *employed* fathers (336) should lose authority (cell 4).

$$\frac{420}{1000} \times 800 = 336$$

To determine the expected frequency in any cell under the assumption of independence, we multiply a cell's row margin total by its column margin total and divide by the total sample size.

Employment Status and Change in Authority: Expected Distribution of 1,000 Fathers under the Assumption of Independence between the Attributes

(Hypothetical data)

	Father Unemployed	*Father Employed*	
Father retains authority	116	464	580
Father loses authority	84	336	420
	200	800	1,000

This 4-fold table gives the *expected distribution* under the assumption of independence between employment status and authority in the family.

Can we reject the hypothesis that there is no association between a father's employment status and his authority in the family? A chi-square table can be used to test the significance of the difference between the *observed* frequency distribution and the frequency distribution *expected* under the hypothesis of independence between employment and authority. We have already spoken of the normal and the t distribution. Chi-square is the name of another distribution. It has been found that the difference between observed and expected frequency distributions is distributed according to the chi-square distribution if the variables are actually not associated. Chi-square is defined as

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e} \tag{22}$$

where f_o = frequency observed (in sample)

f_e = frequency expected (under hypothesis)

Note that χ^2 is always positive (the numerator is squared); that it is small if the differences between the observed and expected frequencies $(f_o - f_e)$ are small; and that χ^2 can be used where there are more than two proportions to be compared.

The computation of χ^2 for our problem is given in Table 7-2.

Table 7-2. Computation of Chi-Square to Test for Independence between Employment Status and Change in Authority: Sample of 1,000 Fathers
(Hypothetical data)

Cell	Frequency		*Difference between Observed and Expected*	*Difference between Observed and Expected Is Squared*	*Chi-Square Ratio*
	Observed (f_o)	*Expected* (f_e)	$(f_o - f_e)$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
Unemployed Father:					
Retains authority	100	116	−16	256	2.207
Loses authority	100	84	16	256	3.048
Employed Father:					
Retains authority	480	464	16	256	.552
Loses authority	320	336	−16	256	.762
Total:	1,000	1,000	0		$\chi^2 = 6.569$

Before we can interpret χ^2, we must know the number of degrees of freedom in our problem, since there are different sampling distri-

butions of χ^2 for different degrees of freedom.[3] Here, "degrees of freedom," is the freedom we would have in filling up the cell frequencies, subject to the requirement that the cell frequencies add up to the margin totals. If the margin totals are given, and we place 116 in the first cell, there is no freedom left to fill the other cells; the frequencies are already determined. Consequently, there is only one degree of freedom in this 2×2 contingency table.

116		580
		420
200	800	1,000

If a table has s rows and t columns, and the row and column totals are fixed, the number of degrees of freedom are $(s - 1)(t - 1)$. A four-fold table usually has $(2 - 1)(2 - 1)$, or one degree of freedom, and a 3×4 table has 2×3, or 6 degrees of freedom.

According to the calculations of Table 7-2, chi-square equals 6.569. With one degree of freedom, the probability of getting so large a sample chi-square from a universe with no association between employment status and paternal authority is less than 2 per cent (see page 172). Therefore at the 5 per cent significance level, we reject the hypothesis of no association between the father's employment status and his authority in the family. These are the same results that we got when we used the z-test for difference between proportions and applied the normal curve. When the number of degrees of freedom is 1, the sampling distribution of the square root of a variable distributed by the chi-square distribution is distributed by the normal distribution.

The sampling distributions of chi-square for different degrees of freedom all begin at zero and extend indefinitely in a positive direction. The skewness decreases and approaches a normal curve as the number of degrees of freedom increases. The sampling distribution of chi-square is given in Fig. 7.10 for 6 degrees of freedom.

[3] If we change the *mean* or the *variance* of a normal population, we get a normal curve which is located differently, and shaped differently. Similarly, if we change what is called the *degrees of freedom* of the chi-squared distribution, we get a chi-square curve which is shaped differently.

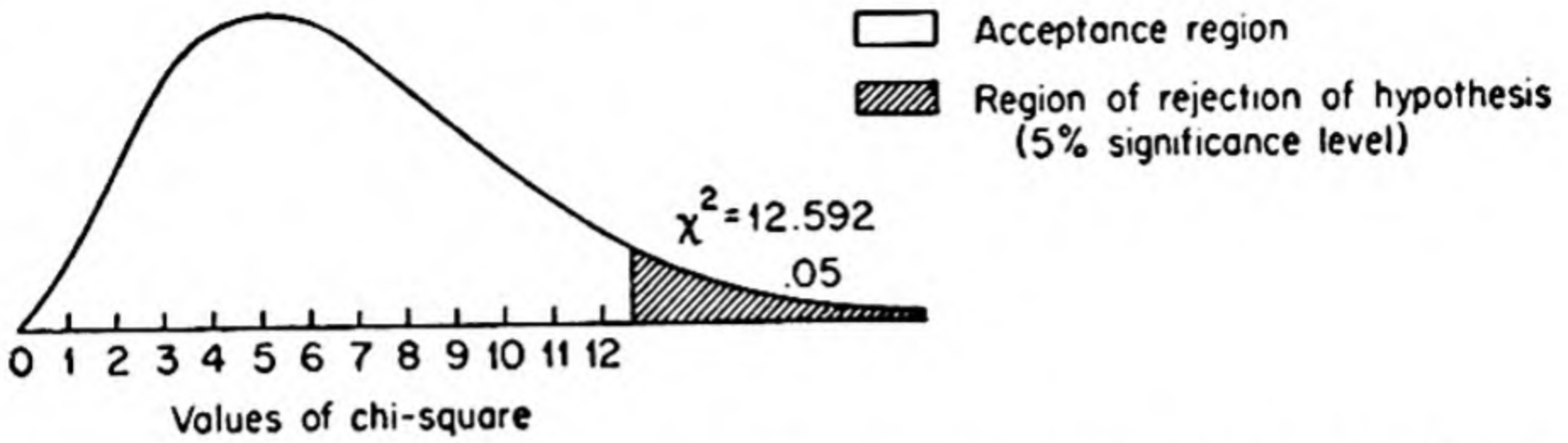


Fig. 7.10. Random sampling distribution of chi-square for six degrees of freedom.

Some Significant Features of Chi-square. A chi-square test can be used in problems with more than four cells. A problem involving only four cells was used here to show that the chi-square test and the z-test of significance give the same result where there is only one degree of freedom.

To apply the chi-square test, every *expected* cell frequency should be at least 5.[4] Several cells may be combined, if necessary, to give the desired size. The fact that intervals may be combined in different ways can alter the results.

The chi-square test ignores the signs of the deviation $(f_o - f_e)^2$. Deviations must be examined for significant peculiarities. In problems where we have more than four cells, it is especially important to examine individual cell deviations.

Percentages must not be used in the distribution of chi-square unless a correction is made for the sample size. Percentages imply a sample size of 100.

The number of degrees of freedom is determined by the number of cells and not by the number of cases.[5]

KEY TERMS

chi-square
contingency table
degrees of freedom
independence
margin totals
observed frequency and expected frequency
sampling distribution of chi-square

[4] A correction factor can be used in computing chi-square from a four-fold table when the expected cell frequencies are small—every deviation of observed from expected cell frequency is reduced by one-half unit.

[5] There are many ways of testing association in contingency tables. The chi-square test is the oldest, and most widely used, but it is not the best for every purpose. Particularly when the cell entries are small, other tests will be more appropriate, e.g., the Fisher exact test and the maximum likelihood test.

REFERENCES

Treloar, Alan E., *Biometric Analysis*, chaps. 14 and 19. Minneapolis: Burgess Publishing Co., 1951.

Freund, John E., *Modern Elementary Statistics*, chap. 15. New York: Prentice-Hall, Inc., 1952.

Peatman, John G., *Descriptive and Sampling Statistics*, chap. 15. New York: Harper and Bros., 1947.

Snedecor, George W., *Statistical Methods*, 4th ed., chaps. 1 and 9. Ames, Iowa: The Iowa State College Press, 1946.

EXERCISES

1. At a certain university, a random sample of 400 men and 300 women were polled on the question: Do you think the legal voting age should be changed from 21 to 18? Among the men: 200 said yes, 160 said no, and 40 were undecided. Among the women: 150 said yes, 140 said no, and 10 were undecided. Is there any association between sex and opinions on the question? (*Note:* in computing chi-square, use all of the information given.)

2. A study of TV-listener habits in three occupational categories shows the following results:

	Professional	*Skilled Worker*	*Unskilled Worker*	
Listen regularly	20	100	80	200
Listen occasionally	40	60	60	160
Do not listen	40	40	60	140
	100	200	200	500

(a) Assuming independence between occupation and TV-listening habits, what are the expected frequencies?

(b) What is the chi-square?

(c) What is the probability of this chi-square? What conclusions can be drawn?

CHAPTER 8

MEASURES OF ASSOCIATION

8.1. Association of Discrete Variables

Four-Fold Table Analysis. Let us re-examine the relationship between unemployment and loss of authority in the Detroit sample of Chapter 7. Table 8-1 gives the employment status of the sample of 1,000 fathers; Table 8-2 gives their loss or retention of authority during an economic recession.

Table 8-1. Employment Status of Sample of 1,000 Fathers
(Hypothetical data)

	Number	*Per Cent*
Father unemployed:	200	20%
Father employed:	800	80%
Total:	1,000	100%

Table 8-2. Change in Authority during Recession: 1,000 Fathers
(Hypothetical data)

	Number	*Per Cent*
Father retains authority:	580	58%
Father loses authority:	420	42%
Total:	1,000	100%

Twenty per cent of the sample of fathers are unemployed and 42 per cent lose authority. Are fathers who are unemployed more likely to lose authority than employed fathers? To get at the association between employment status and authority, we cross-tabulate, i.e., we fill in the cells of the four-fold table, using the information in Tables 8-1 and 8-2 as marginal totals.

Table 8-3 shows that 40 per cent of employed fathers and 50 per cent of unemployed fathers lose authority. Note that we use as the base of the percentages the characteristic we regard as the causal factor—here, employment status. A chi-square of 6.569, with one degree of freedom, signifies that the 10 percentage points difference in loss of authority between employed and unemployed fathers could have occurred in less than two samples out of 100 if there were no association between authority and employment status. (See page 126.) Consequently, we reject the hypothesis of no association, and

Table 8-3. Employment Status and Change in Authority: 1,000 Fathers
(Hypothetical data)

	Number			Per Cent		
	Father Unemployed	*Father Employed*		*Father Unemployed*	*Father Employed*	
Father retains authority	100	480	580	50%	60%	58%
Father loses authority	100	320	420	50%	40%	42%
Total:	200	800	1,000	100%	100%	100%

we say that unemployed fathers are more likely to lose authority in their family than are employed fathers.

Unemployment and loss of authority appear to be associated. Could this association perhaps have come about because both factors are related to a third factor, nativity? Are foreign-born fathers generally the first to lose employment and also the first to lose authority in their family?

We can control the influence of nativity by subdividing the original four-fold table into two tables, one for native-born fathers, the other for foreign-born.

Table 8-4. Employment Status and Change in Authority by Nativity: 1,000 Fathers
(Hypothetical data)

	Native-Born Fathers			Foreign-Born Fathers			
	Unemployed	*Employed*			*Unemployed*	*Employed*	
Father retains authority	48	352	400	Retains	52	128	180
Father loses authority	27	173	200	Loses	73	147	220
	75	525	600		125	275	400

In each subsample, the two original factors, employment status and authority, appear to be independent. The proportion of unemployed fathers who lose authority is almost the same as the proportion of employed fathers who lose authority within each nativity group.

(We could actually test for independence by the use of chi-square.) The fact that within each nativity group there is no association between employment and authority would tend to indicate that there is no direct causal association between employment status and family authority, but rather a spurious relationship between these two characteristics. A *spurious* relationship, either partial or complete, can be shown if there is an antecedent variable related to both original characteristics (employment status and family authority), which, in subdivision, reduces (or removes) the original relationship. With nativity as the antecedent variable, we have all the prerequisites for spuriousness.

(1) Nativity is an antecedent variable; it preceded in time both employment status and authority.

(2) Nativity is related to both employment status and authority. A foreign-born father is more likely to be unemployed than a native-born father; he is also more likely to lose authority within his family.

(3) If we control, or remove the influence of nativity through subdividing the original four-fold table, within each nativity group the original association between employment status and authority of father is reduced, or disappears.

Another factor that may be related to employment status and change in family authority is excessive drinking. Could the association between unemployment and loss of authority have come about because both factors are related to excessive drinking? Table 8-5 gives the subdivision by excessive drinkers and moderate or non-drinkers.

Table 8-5. Employment Status and Change in Authority by Drinking Habits: 1,000 Fathers
(Hypothetical data)

	Excessive Drinkers			Moderate or Nondrinkers		
	Unemployed	*Employed*		*Unemployed*	*Employed*	
Father retains authority	38	82	120	62	398	460
Father loses authority	62	118	180	38	202	240
	100	200	300	100	600	700

Again the relationship between employment status and loss of authority is reduced by the introduction of this third relevant factor. Within each subsample, the proportion of unemployed fathers who lose authority is almost the same as the proportion of employed fathers who lose authority. We could again test for the independence of these two characteristics through chi-square.

But we cannot say, on the basis of this third factor, that the original relationship is a spurious one, and that there is no direct causal association between employment status and authority. Drinking is not necessarily an antecedent variable; it is quite possibly an intervening one. Unemployment can lead to excessive drinking, which in turn can lead to loss of authority. When a third *intervening* variable, related to both original variables, is introduced, we cannot conclude that there is no direct causal association between two original factors, even though the original association disappears in each subsample.

When we want to control factors through subdivision, a large number of cases are required to fill every cell, and a great deal of information is needed with regard to every case. For example, if we want to control the influence of *nativity*, with its two categories: native- and foreign-born, and *education*, with four categories: completed college, high school, grammar school, less than eighth grade, at the same time on the employment-authority relationship, each with two categories, we shall have 32 cells ($2 \times 4 \times 2 \times 2$).

In addition to requiring a vast number of cases to fill all the cells, further and further subdivision can soon reach a point of diminishing returns. For example, if we were to subdivide the population of the United States in order to predict success or failure in marriage, how many people would there be who have been married to a person of another religion 5 to 9 years, have three children, no in-laws within a radius of fifty miles, and a stable income of between $5,000 and $10,000 during this period? Minute subclassification can lose any predictive value, since there is almost no universe to which it applies.

We know that if an initial relationship between two variables does not disappear after subdivision by every relevant antecedent variable, the initial relationship is a causal one. But how do we know when all relevant variables are exhausted? Such knowledge can be gained through a controlled experiment and not through continuous subdivision. We can try to get at relevant antecedent variables in subdivision by making intensive studies of those cases that seem to

Number of Cells Required to Relate Employment Status and Change in Father's Authority, Controlling Nativity and Education Through Subdivision

COMPLETED COLLEGE

	Foreign-Born		*Native-Born*	
	Unemployed	Employed	Unemployed	Employed
Retain				
Lose				

COMPLETED HIGH SCHOOL

	Foreign-Born		*Native-Born*	
	Unemployed	Employed	Unemployed	Employed
Retain				
Lose				

COMPLETED GRAMMAR SCHOOL

	Foreign-Born		*Native-Born*	
	Unemployed	Employed	Unemployed	Employed
Retain				
Lose				

LESS THAN 8 YEARS OF SCHOOLING

	Foreign-Born		*Native-Born*	
	Unemployed	Employed	Unemployed	Employed
Retain				
Lose				

deviate from the general trend. If, for example, 95 per cent of unemployed foreign-born fathers with less than eight years of schooling lost authority in their family, what characteristics of the other 5 per cent helped them to retain authority?

A Measure of the Degree of Association. In our original four-fold table we found the following relationship between employment status and authority in the family:

	Unemployed Father	*Employed Father*		
Father retains authority	100	480	580	50% of unemployed fathers lose authority
Father loses authority	100	320	420	40% of employed fathers lose authority
	200	800	1,000	

The chi-square test of significance indicates that this is not a chance association due to fluctuations in random sampling. We want to measure the degree of association.

There are many different measures of the degree of association for discrete variables, based on different assumptions and having different properties.[1] The formula for one frequently used measure of the degree of association is the square root of T^2 where

$$T^2 = \frac{\chi^2}{n\sqrt{(t-1)(s-1)}} \tag{23}$$

and where s equals the number of rows, and t, the number of columns. For a four-fold table, $T^2 = \chi^2/n$, since $(s-1)(t-1) = 1$.

The association between employment status and authority, as measured by T, is equal to 08, not a very high association.

$$T^2 = \frac{\chi^2}{n} = \frac{6.569}{1000} = .006569$$

$$T \doteq .08$$

The T would equal 1 if the association between the characteristics were perfect, and the distribution had the same row and column totals—here, 580 and 420.

[1] For an analysis of such measures, see the texts by Guilford and Peters and Van Voorhis cited on page 135.

0	580	580
420	0	420
420	580	1,000

$$T^2 = \frac{\chi^2}{n} = \frac{1000}{1000} = 1$$

$$\therefore T = 1$$

If the two variables are independent, the frequency observed equals the frequency expected, and chi-square will equal 0. Hence T will equal 0.

116	464	580
84	336	420
200	800	1,000

$$T^2 = \frac{\chi^2}{n} = \frac{0}{1000} = 0$$

$$\therefore T = 0$$

The highest T that a four-fold table can have where the margin totals are not symmetrical but are fixed at (580 : 420) and (800 : 200) is about .59.

REFERENCES

Guilford, J. P., *Fundamental Statistics in Psychology and Education*, 2nd ed., chap. 13. New York: McGraw-Hill Book Company, Inc., 1950.

Hagood, M. J., and D. O. Price, *Statistics for Sociologists*, chap. 21. New York: Henry Holt and Company, 1952.

Kendall, P. L., and P. F. Lazarsfeld, "Problems of Survey Analysis" in *Continuities in Social Research*, edit. Robert Merton and Paul Lazarsfeld. Glencoe, Illinois: The Free Press, 1950.

Peters, C. C., and Van Voorhis, W. R., *Statistical Procedures and their Mathematical Bases*, chap. 13. New York: McGraw-Hill Book Company, Inc., 1940.

Yule, G. U., and Kendall, M. G., *An Introduction to the Theory of Statistics*, 14th ed., pp. 50–65. New York: Hafner Publishing Company, Inc., 1950.

EXERCISES

1. Given the relationship between home ownership and race in the accompanying four-fold table (numbers are in millions of people):

	Negro	*White*	*Million*
Home Owner	.5	19.5	20.0
Renter	2.5	17.5	20.0
	3.0	37.0	40.0

(a) What is the probability:

(1) That a Negro is a home-owner?
(2) That a White is a home-owner?
(3) That a home-owner is a Negro?
(4) That a home-owner is a White?

(b) If we were to choose a Negro at random, would we estimate that he is a home-owner or a renter? What measure of central value are we using?

2. In a study on the loyalty oath made at a California university, students were asked whether (1) university employees should be compelled to sign a loyalty oath and (2) nonsigners should be dismissed. The results:

		Question 1	
		Yes	No
Question 2	Yes	106	3
	No	50	207

109		
257		
156	210	366

(a) If Questions 1 and 2 are indicative of the same basic attitude, in what cells of the contingency table would you expect all the cases to lie?

(b) If you wanted to follow up the results of this study with some intensive interviews, on what cases would you concentrate? (*Source:* D. Wilner and F. Fearing, "The Structure of Opinion: A Loyalty Oath Poll," *The Public Opinion Quarterly*, Winter 1950–51, pp. 729–43.)

3. Lazarsfeld and Kendall attempted to find out how Frederic Wakeman's novel, *The Hucksters*, and the subsequent motion picture based on the novel, might have affected attitudes toward radio advertising. (Wakeman gives a satirical indictment of radio advertising.) The Lazarsfeld-Kendall findings are presented in the table.

Percentage with Highest Criticism Score* According to Exposure to *The Hucksters* and Education

	College	*High School*†
Read *The Hucksters:*	49%††	35%
Did not read it:	35	23
Saw *The Hucksters:*	48	31
Did not see it:	36	23

SOURCE: P. Lazarsfeld and P. Kendall, *Radio Listening in America* (New York: Prentice-Hall, Inc., 1948), p. 78.

* The higher the criticism score, the more negative the attitude toward commercials.

† Respondents with only grade school education have not been included in the analysis. So few of them had either read *The Hucksters* or seen the movie that it was impossible to study their attitudes toward radio commercials in relation to their exposure.

†† The sample size is not specified.

(a) Is there a relation between exposure to Wakeman's thesis (either through reading the book or seeing the movie) and attitude toward commercials?

(b) Does this relation hold only because college-educated people are more critical, and therefore more likely to expose themselves to *The Hucksters*? What happens when education is controlled?

4. The accompanying graphs are based on a public opinion survey of 1,100 representative adult white men in the United States, undertaken in the summer of 1945. (*Source:* Richard Centers, *The Psychology of Social Classes.* Princeton: Princeton University Press, 1949).

Figure 8.1 gives the relationship between present *occupational stratum* and *class identification.*

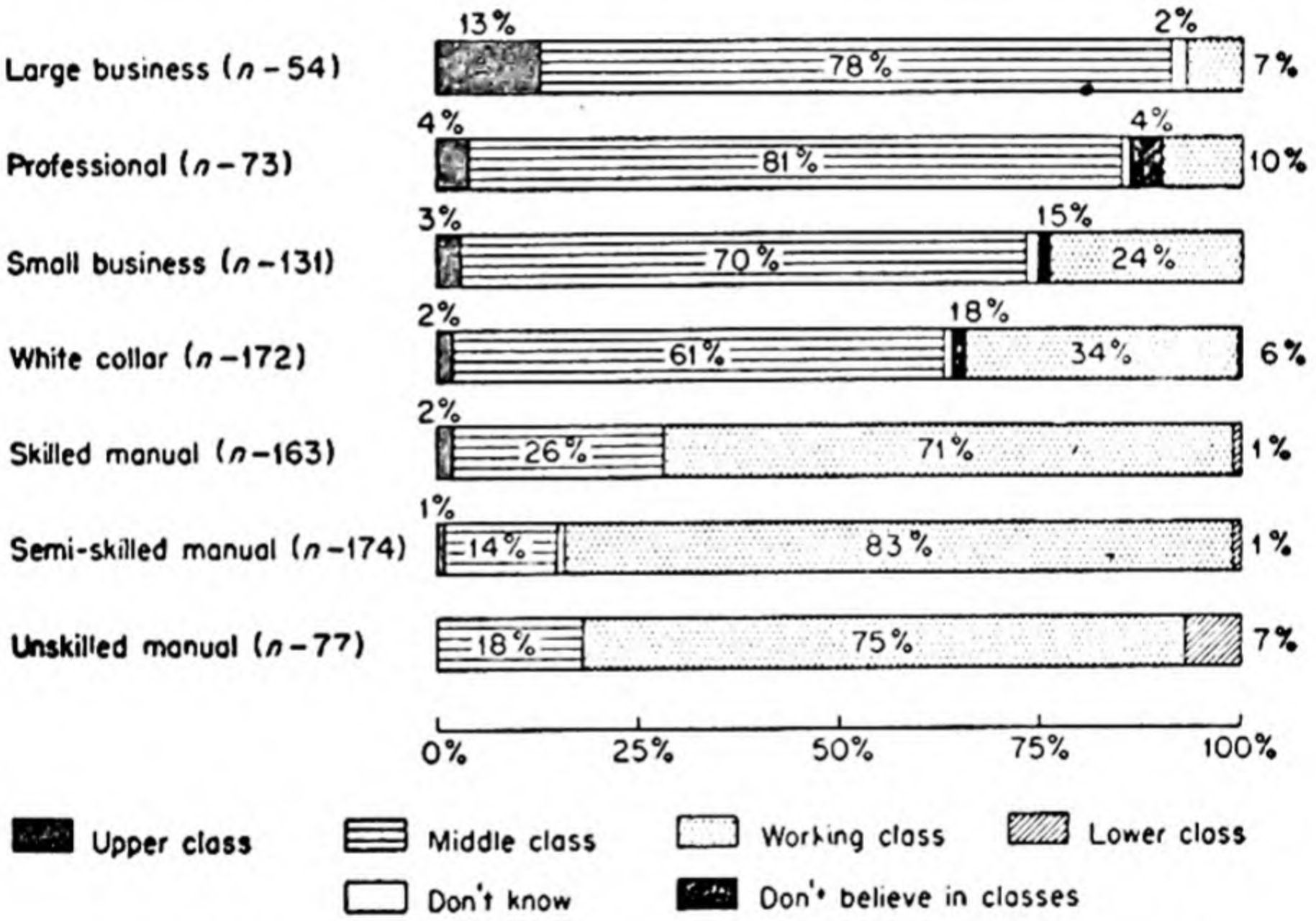


Fig. 8.1. Class identifications of urban occupational strata.

Question 1. What is the difference in class identification between business, professional, and white-collar workers as compared with manual workers? If the occupational strata represent an occupational hierarchy, where does the break in identification occur?

Question 2. Figure 8.2 gives the relationship between class identification and politico-economic attitude. What is the relationship between conservatism-radicalism and middle and working class identification?

Question 3. One of the reasons why class identification and conservatism-radicalism may not be even more closely related to each other is probably because there are variables significantly related to

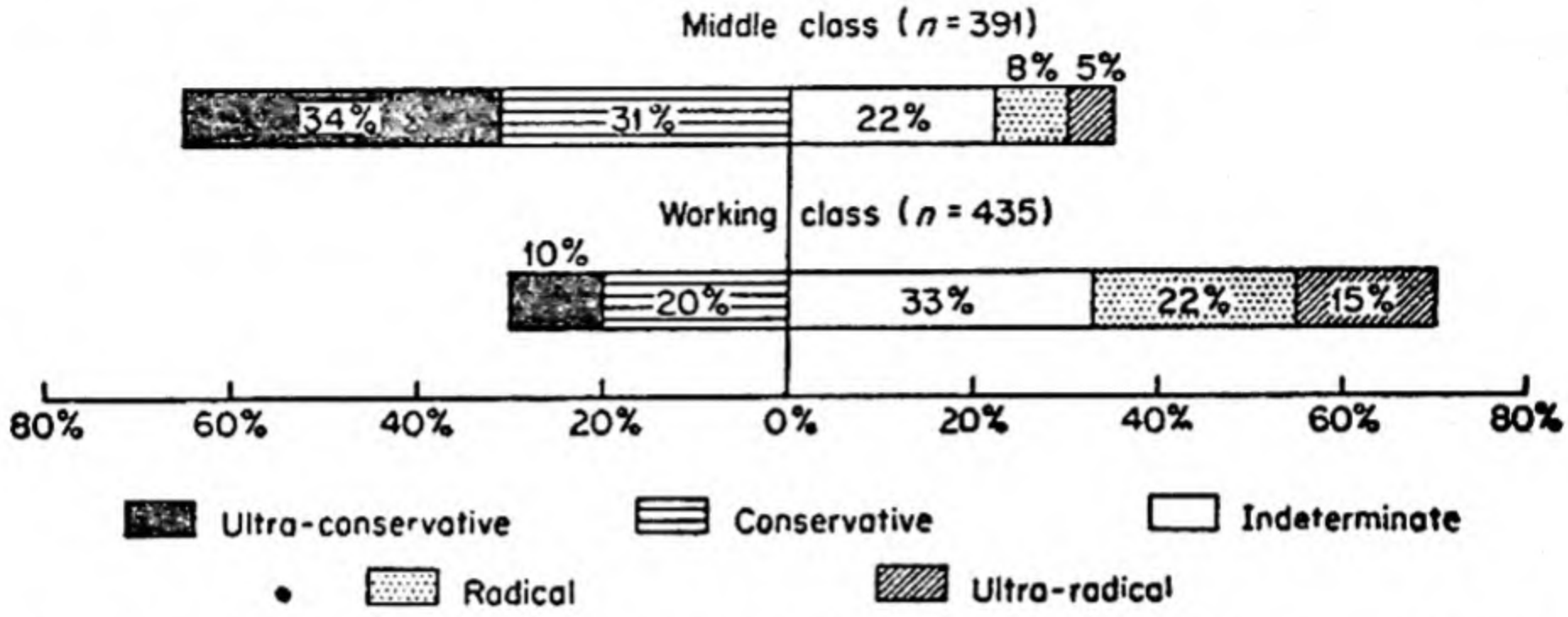


Fig. 8.2. Class differences in conservatism-radicalism: urban population.

class identification, but not to conservatism (or vice versa). One such variable is education. Speculate on the influence of this variable, or see Centers, *Psychology of Social Classes*, pp. 203–4, 94–5.

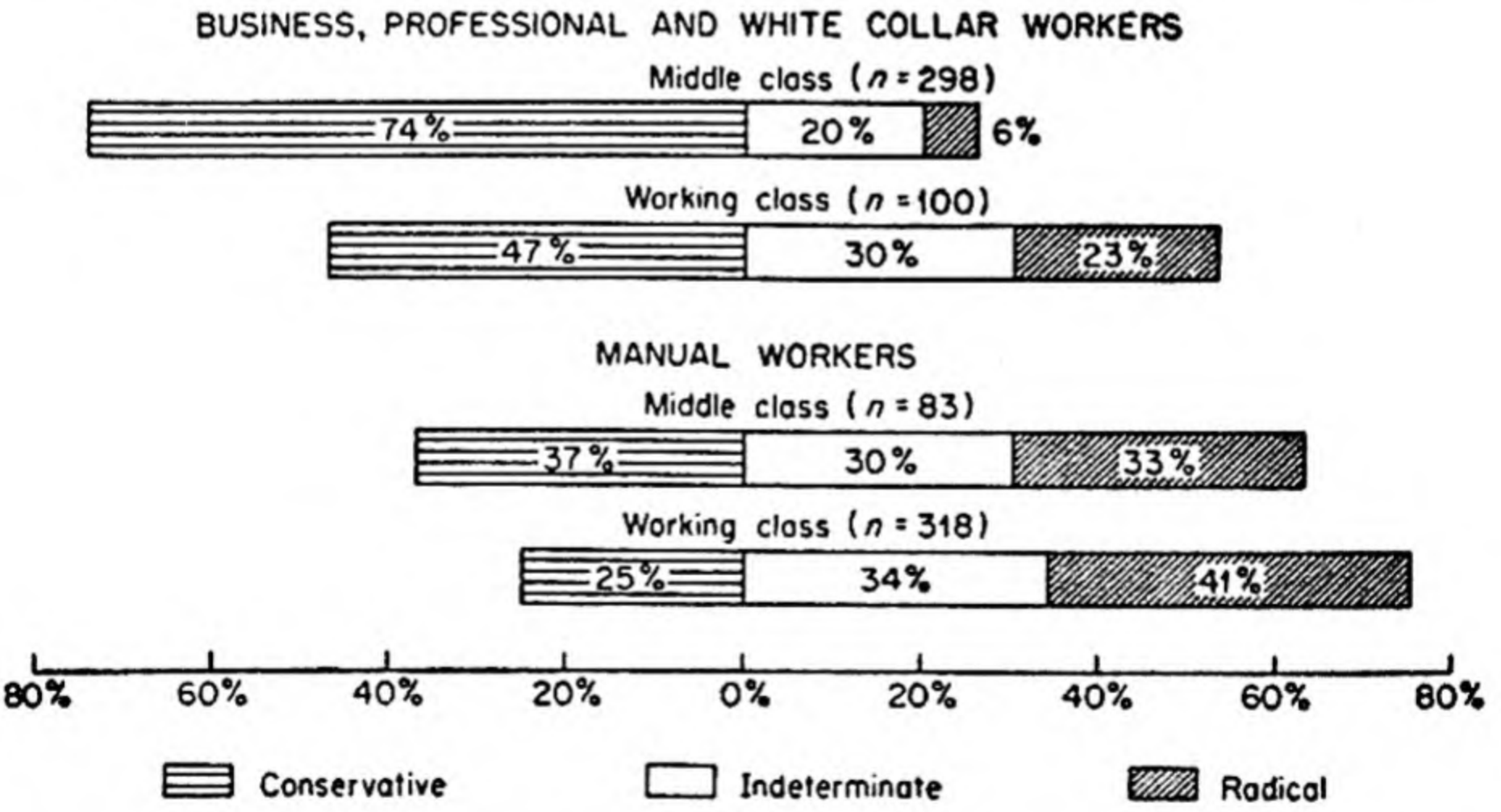


Fig. 8.3. Stratum and class differences in conservatism-radicalism: urban population.

Three variables have been introduced: occupational stratum, class identification, and politico-economic attitude. Centers hypothe-

sizes that people in the same broad occupational stratum have different class identifications (Fig. 8.1) because they differ in attitude or politico-economic orientation from the members of their own occupational stratum. To test this hypothesis, he relates class identification and attitudes, holding constant membership in occupational stratum (Fig. 8.3).

Question 4. Is the hypothesis confirmed? On the basis of the data in Fig. 8.3, what factors appear to recruit people to social classes?

Question 5. Most of the interviewers were themselves members of the middle class. In view of the fact that responses reported by working-class interviewers on working people tend to be more radical than those reported by middle-class interviewers on working people, how might the results be biased?

8.2. Association of Continuous Variables

Linear Regression. Problem. We want to determine the relationship between a child's frustration because of disciplinary action at home and his subsequent behavior at school.

Procedure. Our sample consists of 5 four-year-old children selected at random from a nursery school. There are 2 measurements for each child: an X value, which is a rating on a scale measuring frustration with parents over matters of discipline, and a Y value, which is a rating on a scale measuring assertiveness and aggression in the class-

Table 8-6. Frustration Rating and Rating in Classroom Aggression for Five Nursery-School Children
(Hypothetical data)

Child	*Frustration Rating* (X)	*Rating in Classroom Aggression* (Y)	*Deviations from Mean Rate in Classroom Aggression* ($Y - \bar{Y} = y$)	*Squares of Deviations from Mean Rate in Classroom Aggression* ($(Y - \bar{Y})^2 = y^2$)
1	3	4	−3	9
2	6	6	−1	1
3	4	7	0	0
4	7	9	2	4
5	10	9	2	4
Sum:	30	35	0	18
Mean:	6	7		

Variance in Rating in Classroom Aggression: $\sigma_y^2 = \frac{\Sigma y^2}{n} = \frac{18}{5} = 3.6$

room. Each scale extends from 0 to 10, the higher numbers representing greater frustration and aggression.

We do not apply regression and correlation techniques to a sample this small, but we shall do so here for ease in tabulation. Given in Table 8-6 are the frustration rating X and the rating in classroom aggression Y for each of the 5 children.

If we want to estimate a sixth nursery-school child's rating in classroom aggression, we would use our mean aggression rating as the best estimate. The mean is 7, and the variance about the mean is 3.6.

But if there is a known relationship between the frustration-with-parent-rating and the rating in classroom aggression, we can predict a child's classroom aggression with greater accuracy from his own frustration rating than from the mean aggression rating of all the children.

We want to know how aggression ratings vary with ratings in frustration. Our first step is to draw a graph with frustration rating, regarded as the independent variable, measured along the horizontal axis (called the X axis), and rating in classroom aggression, regarded as the dependent variable, measured along the vertical axis (called the Y axis). The graph, called a *scatter diagram*, has five points plotted on it, each point representing a child's score on the two scales. The first child, for example, has a rating of 3 in the frustration scale and 4 in the aggression scale. Consequently, his point is plotted above $X = 3$ and across from $Y = 4$; it can be written as $X = 3$, $Y = 4$, or 3,4. The scatter diagram should indicate whether the points appear to cluster around a straight line which could be drawn to fit the data.

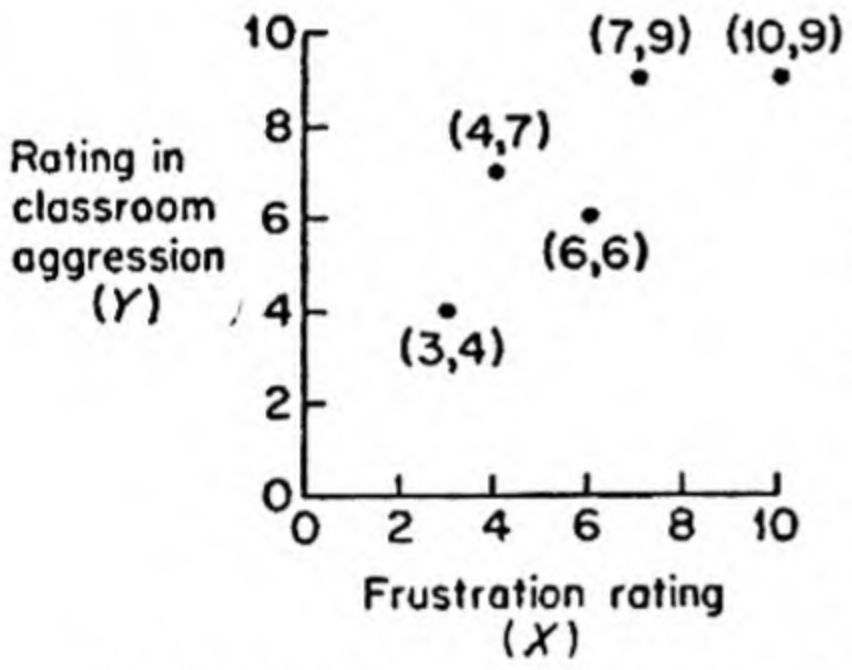


Fig. 8.4. Scatter diagram of the relationship between rating in classroom aggression and frustration rating. (Hypothetical sample of five nursery school children.)

The relationship between the frustration rating and the rating in classroom aggression does appear to be linear. Aggression tends to increase at a constant rate with an increase in frustration. We could draw free-hand any number of good straight lines to fit the data. We

standardize our procedure by getting a theoretical line of best fit, a line which will best represent the general trend of the observed data. This line of best fit, also called the regression line of Y on X, is calculated mathematically as the line from which the sum of the squares of the vertical deviations in observed values are smaller than the sum of the squares of deviations from any other line.[2]

If the *observed values* of the rating in classroom aggression are symbolized by Y, and the *estimated values* on the line of regression by Y_c (Y computed), then the vertical deviation of observed from estimated

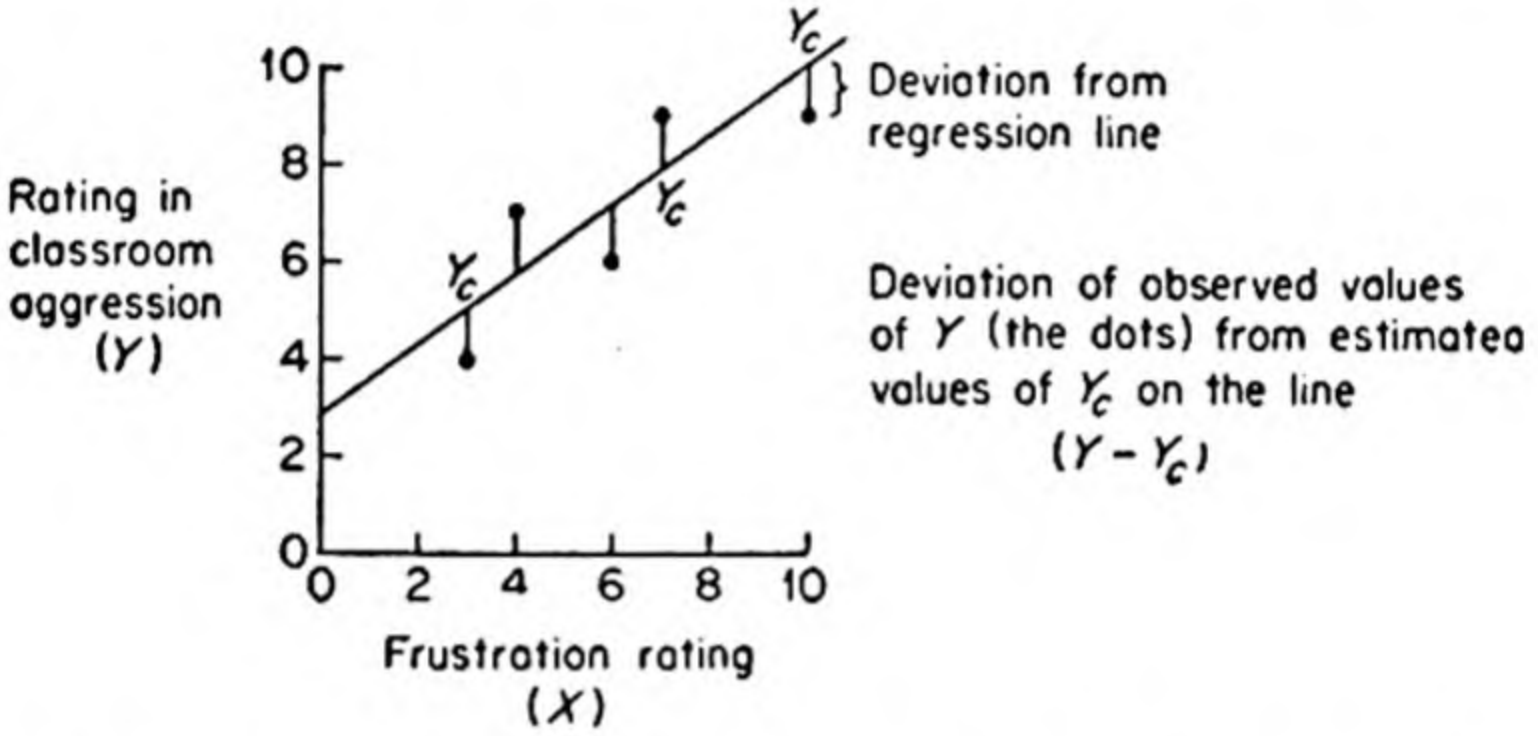


Fig. 8.5. Linear relationship between rating in classroom aggression (Y) and frustration rating (X). (Hypothetical sample of five nursery school children.)

value, the *error in estimate*, is $Y - Y_c$, and the sum of the vertical deviations, all the errors of estimate, is $\Sigma(Y - Y_c)$. Since positive deviations can cancel out negative deviations, we calculate a line in which the sum of the *square* of the deviations, and not the sum of the deviations, is made a minimum.

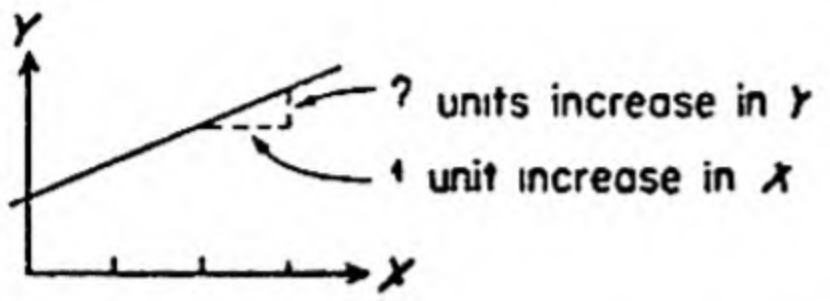
The equation for any nonvertical straight-line relationship between an independent variable X and a dependent variable Y is $Y = a + bX$.

The Y is the dependent variable and the X is the independent variable. We shall want to predict the variable Y, given fixed values

[2] The term "regression" comes from Galton and Pearson. In his studies of inheritance, Pearson shows that although tall fathers tend to have tall sons, the sons tend to be shorter than their fathers. There is a regression of sons' heights toward the average height of the population.

of X. The Y is a function of X; with each unit change in X, Y changes b units.

The constants a and b are determined from our observed data; a is the height of the line (Y) when X equals zero, and b is the slope of the line. As X increases one unit, Y changes b units. We shall give two simple practice examples of straight-line equations.



(1) $Y = 2X$ $a = 0$ The a is the height of the line (Y) when X equals zero. In the equation $Y = 2X$, when $X = 0$, $Y = 0$. $[Y = 2(0) = 0]$.

$b = 2$ As X increases one unit, Y changes b units.

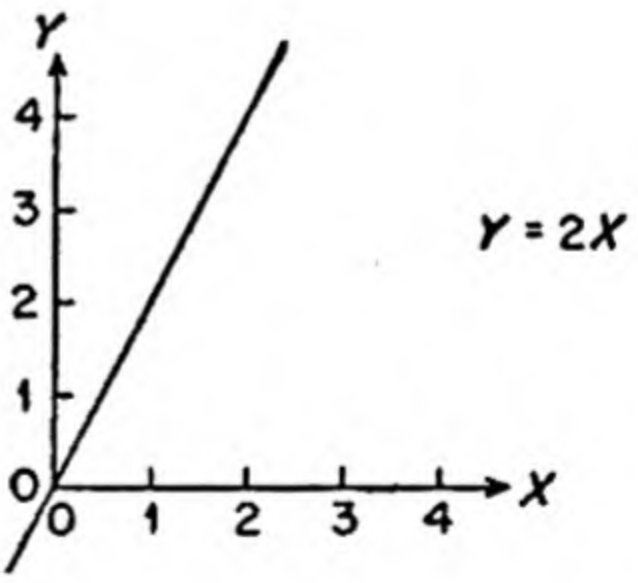
In the equation $Y = 2X$, Y increases 2 units for every unit increase in X. $[Y = 2(1) = 2]$.

At least two points are needed to plot a straight line. Three points

Co-ordinates of three points on line $Y = 2X$

X	Y
0	0
2	4
-1	-2

Graph of line $Y = 2X$



are calculated for the line $Y = 2X$, and the line is drawn on the accompanying graph.

(2) $Y = 1 - 1X$ $a = 1$ The a is the height of the line (Y) when X equals zero. In the equation $Y = 1 - 1X$, when $X = 0$, $Y = 1$.

$b = -1$ The b is the slope of the line. As X increases one unit, Y changes b units.

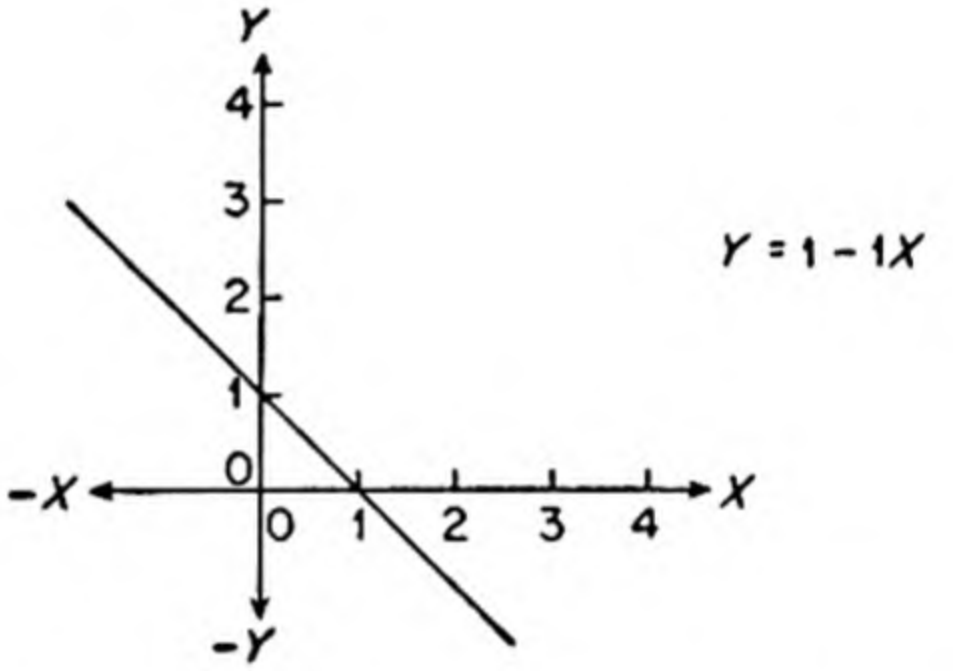
In the equation $Y = 1 - 1X$, Y decreases one unit for every unit increase in X.

In the line of best fit, the constants a and b are chosen to make $\Sigma(Y - Y_c)^2$ a minimum, i.e., to minimize the sum of the squares of

Co-ordinates of three points on line $Y = 1 - 1X$

X	Y
0	1
1	0
2	-1

Graph for line $Y = 1 - 1X$



the errors of estimate. It can be shown that when b is set equal to $\Sigma xy/\Sigma x^2$, and a is set equal to $\overline{Y} - b\overline{X}$, we have a line which passes through the scatter diagram in such a way that the sum of squares of deviations from the line are minimized. Deviations are taken along the vertical. We are minimizing the error of predicting a Y, given an X.[3]

$$b = \frac{\Sigma xy}{\Sigma x^2} \tag{24}$$

$$a = \overline{Y} - b\overline{X} \tag{25}$$

where x = deviation of observed value of X from $\overline{X}$

and y = deviation of observed value of Y from $\overline{Y}$

We want to determine the line of best fit in our problem relating frustration ratings to ratings in classroom aggression. To do this,

[3] For students who know some calculus, there is a fairly simple derivation of these formulas in Paul G. Hoel, *Introduction to Mathematical Statistics* (New York: John Wiley and Sons, Inc., 1947), pp. 79–80.

Table 8-7. Calculation of Regression Line of Rating in Classroom Aggression (Y) on Frustration Rating (X)
(Hypothetical sample of five nursery school children)

Child	*Frustration Rating* (X)	*Rating in Classroom Aggression* (Y)	*Deviations from Mean* (x)	(y)	*Squares of Deviations from Mean* (x^2)	(y^2)	*Product of Deviations from Mean* (xy)	*Y Value on Line* (Y_c)	*Deviations from Line* ($Y - Y_c$)	*Squares of Deviations from Line* $(Y - Y_c)^2$
1	3	4	−3	−3	9	9	9	5.10	−1.10	1.21
2	6	6	0	−1	0	1	0	7.00	−1.00	1.00
3	4	7	−2	0	4	0	0	5.73	1.27	1.61
4	7	9	1	2	1	4	2	7.63	1.37	1.88
5	10	9	4	2	16	4	8	9.53	−0.53	0.28
Sum:	30	35	0	0	30	18	19			5.98
Mean:	6	7								

Constants a and b:

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{19}{30} = \underline{.633}^*$$

$$a = \bar{Y} - b\bar{X} = 7 - .633(6) = 7 - 3.80 = \underline{3.20}$$

Regression Line:

$$Y_c = a + bX = \underline{3.20 + .633X}$$

Variance around Mean Y:

$$\frac{\Sigma(Y - \bar{Y})^2}{n} = \frac{\Sigma y^2}{n} = \frac{18}{5} = \underline{3.60}$$

Variance around Line:

$$\frac{\Sigma(Y - Y_c)^2}{n} = \frac{5.98}{5} = \underline{1.20}$$

* To compute b without calculating deviations from a mean, we can convert small x's into large X's. The equation for b is then:

$$b = \frac{\Sigma XY - \dfrac{(\Sigma X)(\Sigma Y)}{n}}{\Sigma X^2 - \dfrac{(\Sigma X)^2}{n}} = \frac{229 - \dfrac{(30)(35)}{5}}{210 - \dfrac{(30)^2}{5}} = \frac{229 - 210}{210 - 180} = \frac{19}{30} = .633$$

we must solve for the two unknown constants, a and b. Our calculations are found in Table 8-7.

The regression line, $Y_c = a + bX$ becomes $Y_c = 3.20 + .633\,X$, as calculated in Table 8-7 and plotted in Fig. 8.6. One point on the line is $a = 3.20 (Y = 3.20$ when $X = 0)$. Another point is the intersection of the two means (6,7).

Coordinates of points on line $Y_c = 3.20 + .633X$

X	Y
0	3.20
6	7
2	4.47

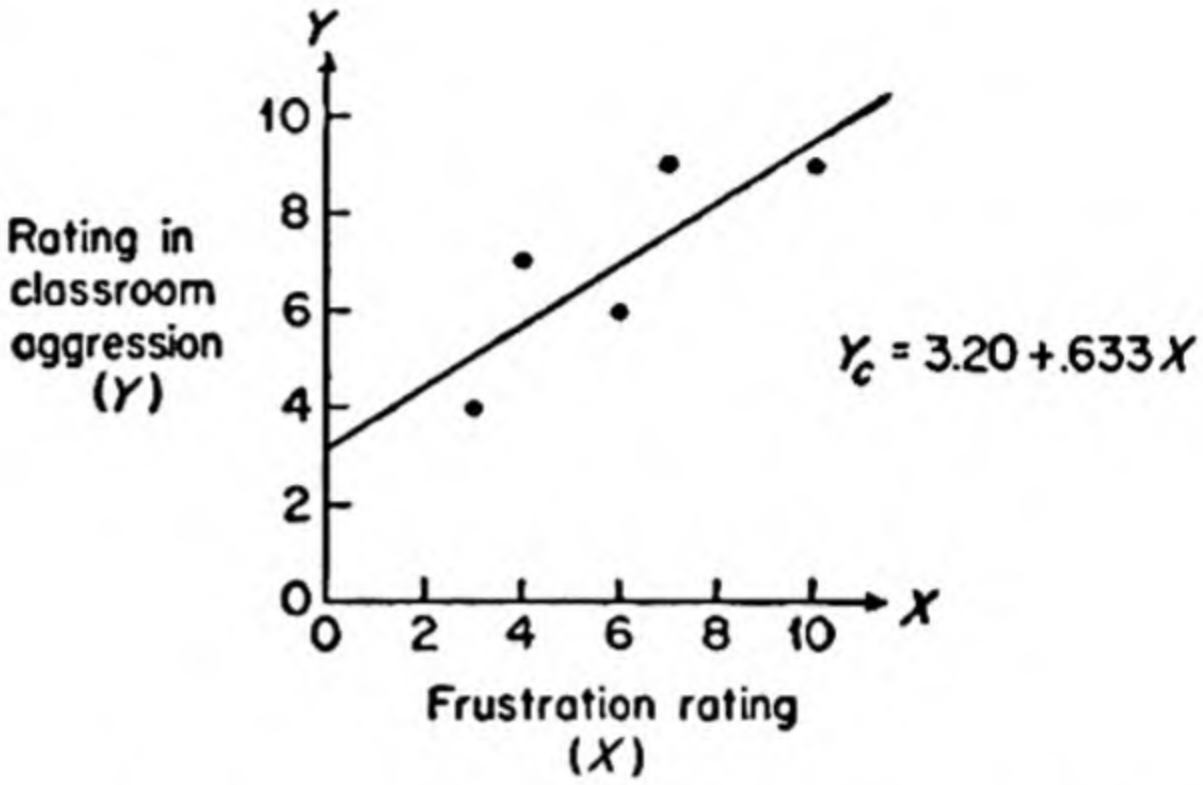


Fig. 8.6. Regression line of rating in classroom aggression (Y) on frustration rating (X). (Hypothetical sample of five nursery school children.)

For a frustration rating of 5 ($X = 5$), the rating in classroom aggression which is on the regression line is 6.37 ($Y_c = 6.37$).

$$Y_c = 3.20 + .633X = 3.20 + .633(5) = 3.20 + 3.17 = 6.37$$

Unless the association between frustration rating and rating in classroom aggression is perfect, the observed values of Y will differ from the estimated values on the line (Y_c), and there will be errors in making an estimate from the line.

We assume that the errors are of the same magnitude all along the line. One measure of the variability of actual values around the line is the standard deviation around the regression line, $\sqrt{\Sigma(Y - Y_c)^2/n}$, called the standard error of estimate:

$$s_{Y \cdot} = \sqrt{\frac{\Sigma(Y - Y_c)^2}{n}} \tag{26}$$

From the standard error of estimate for the sample we can estimate the standard error of estimate for the universe:

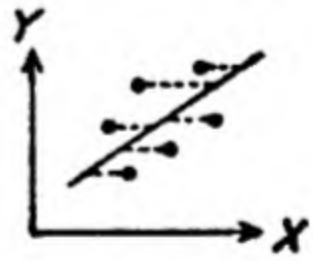
$$\text{est } \sigma_{Y_c} = \sqrt{\frac{\Sigma(Y - Y_c)^2}{n - 2}} \tag{27}$$

If the distribution of the Y values for any given X value in the nursery school universe is *normal*, we can estimate the mean Y value for all children having a particular X value. We can say, for example, that we are 95 per cent confident that the mean aggression rating of all children with a frustration rating of 5 is between two specified points. The 5-children sample upon which our regression line is based is too small to set up confidence limits. As with other sample statistics, we can also perform a test of significance to determine whether or not the variable Y is independent of the variable X.

Note: Unless the actual values fit the line perfectly (i.e., the standard error of estimate equals zero), there are two different lines of regression depending upon which variable is assumed to be fixed or independent. In the regression of Y on X (X considered the independent variable), the squared deviations of the observed Y-values from their computed values are minimized. In the regression of X on Y (Y considered the independent variable), the squared deviations of the observed X values from their computed values are minimized.

$\Sigma(Y - Y_c)^2$ are a minimum
(X is the independent variable)

$\Sigma(X - X_c)^2$ are a minimum
(Y is the independent variable)



REFERENCES

Dixon, W. J., and F. J. Massey, *Introduction to Statistical Analysis*, chap. 11. New York: McGraw-Hill Book Company, Inc., 1951.

Freund, John E., *Modern Elementary Statistics*, chaps. 12 and 13. New York: Prentice-Hall, Inc., 1952.

Hagood, M. J., and D. O. Price, *Statistics for Sociologists*, chap. 23. New York: Henry Holt and Company, 1952.

Snedecor, George W., *Statistical Method*, 4th ed., chap. 6. Ames, Iowa: The Iowa State College Press, 1946.

EXERCISE

Compute the line $Y = a + bX$ from the following data. Plot the line approximately. What is the variance around the mean, the variance around the line? What is the standard error of estimate? (Check your computation of b by using both formulas: the formula with small x's and the one requiring large X's.)

X	Y
1	11
2	9
3	7
4	4
5	3

Linear Correlation. We have observations on frustration rating and rating in classroom aggression for a random sample of five nursery school children. Let us assume that in the universe of nursery school children the distributions of aggression ratings (Y) for given frustration ratings (X) are normal, and, in addition, the distributions of frustration ratings (X) for given aggression ratings (Y) are normal.

We want to know how close is the linear relation between X and Y. We want a measure of the usefulness of the frustration rating (X) in estimating classroom aggression (Y). If we had no knowledge of frustration among the children, the mean aggression rating would be our best estimate of the aggression of any child in the nursery school population to which our sample belongs. The mean aggression rating is a fixed value, the same for each child.

Knowing the frustration rating, the X variable, we hope to improve our estimation of rating in classroom aggression. We measure the improvement by the proportionate reduction in variance around our estimate. The variance around the mean is 3.60. (See Table 8-7.) The variance around the line of regression, the variability still not accounted for even by the relationship of frustration rating to rating in classroom aggression, is 1.20.

The total variance around the mean aggression rating (σ_y^2) can be divided into unexplained variance, the proportion of variance that *cannot* be attributed to variation in frustration rating, and explained variance, the proportion of variance that *can* be attributed to varia-

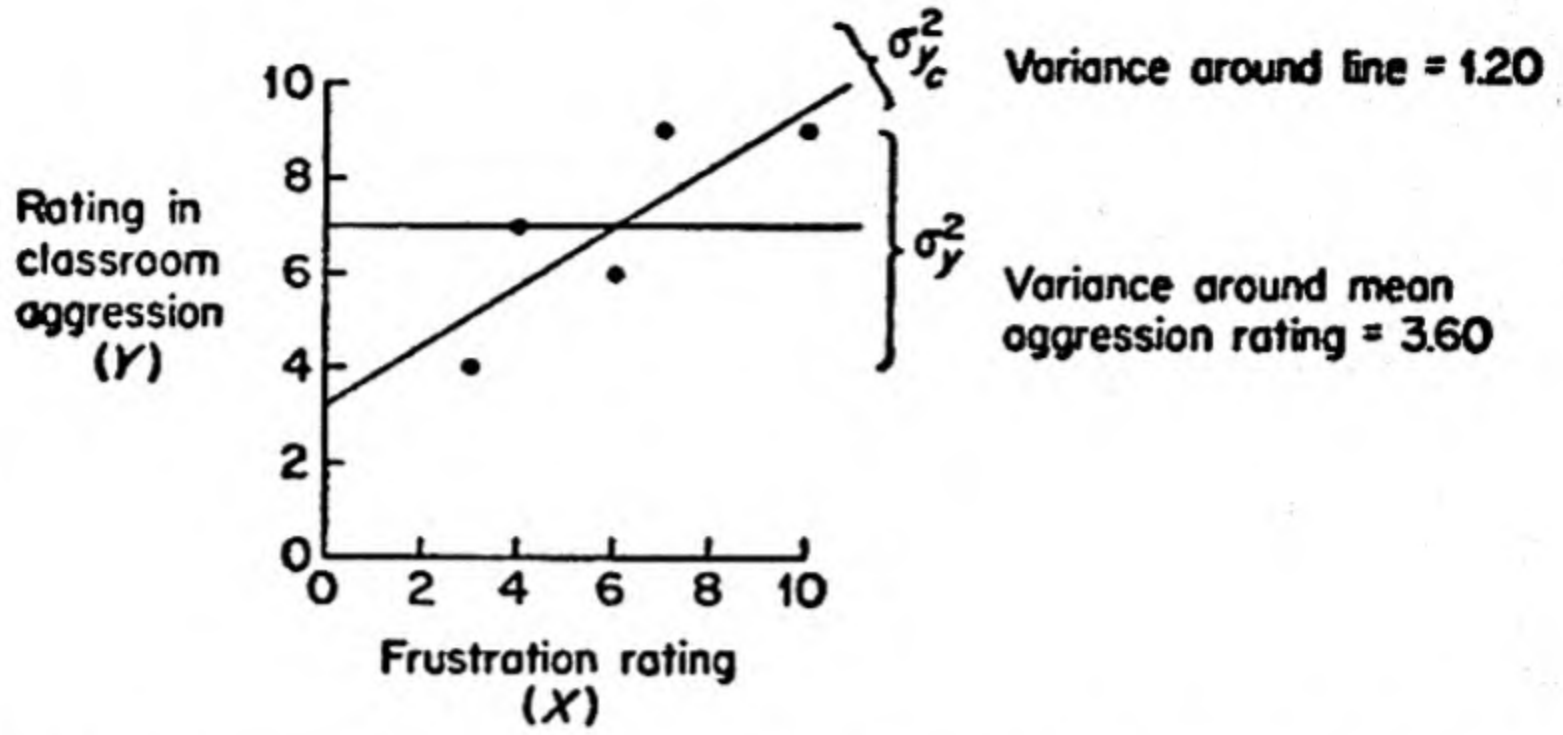


Fig. 8.7. Regression line of rating in classroom aggression (Y) on frustration rating (X) compared with mean rating in classroom aggression. (Hypothetical sample of five nursery school children.)

tion in frustration rating. If the total variance is equal to 1, then:

$$1 = \begin{array}{l}\text{Proportion of variance that can be attributed to variation in frustration rating}\end{array} + \begin{array}{l}\text{Proportion of variance that cannot be attributed to variation in frustration rating}\end{array}$$

The proportion of the total variance that *cannot* be attributed to variation in frustration rating is the variance around the regression line divided by the variance around the mean.

$$\frac{\sigma_{y_c}^2}{\sigma_y^2} = \frac{1.20}{3.60} = .333$$

The proportion of the total variance that *can* be attributed to variation in the frustration rating is $1 - \sigma_{y_c}^2/\sigma_y^2$. It is called the coefficient of determination, or r^2.

$$r^2 = 1 - \frac{\sigma_{y_c}^2}{\sigma_y^2} \quad . \tag{28}$$

$$= 1 - .333 = .667$$

The square root of the coefficient of determination is called the coefficient of correlation.[4]

$$r = \sqrt{1 - \frac{\sigma_{y_c}^2}{\sigma_y^2}} \tag{29}$$

$$= \sqrt{.667} = .82$$

[4] This definition of correlation is a general one, including curvilinear as well as linear correlation.

We can now say (or could say, if our sample were a random one of sufficient size) that there is a positive correlation (r) of .82 between frustration rating and rating in classroom aggression, and that about .67 (r^2) of the total variance around the mean classroom aggression rating can be explained by the relationship with the frustration rating.

The range of the coefficient of correlation goes from -1 through 0 to $+1$. It equals -1 or $+1$ when there is no variance around the regression line; all the observed values are on the line of regression. The association with the X variable accounts for all the variance around mean Y. If we know the frustration rating for any child in the population to which our sample applies, we can estimate perfectly his aggression rating. For each value of X, there corresponds only one value of Y. For $r = +1$, high values of one variable are associated with high values of the other; for $r = -1$, high values of one variable are associated with low values of the other.

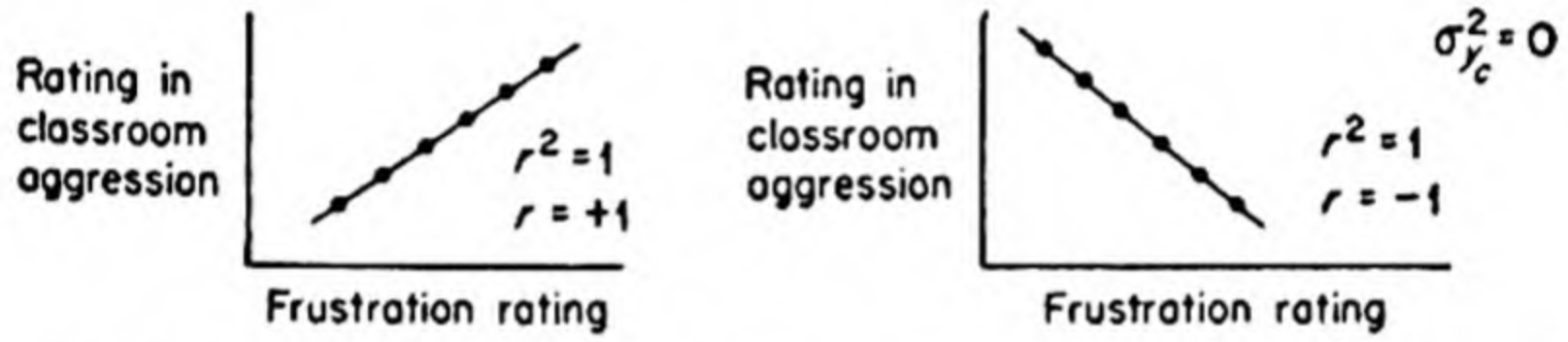


Fig. 8.8. Perfect linear relationship between frustration rating and rating in classroom aggression ($r = +1$ and $r = -1$).

When the variance around the line is as great as the variance around the mean, r^2 and r will equal 0. The regression line is simply a horizontal line through the mean. The frustration rating is no help in estimating rating in classroom aggression.

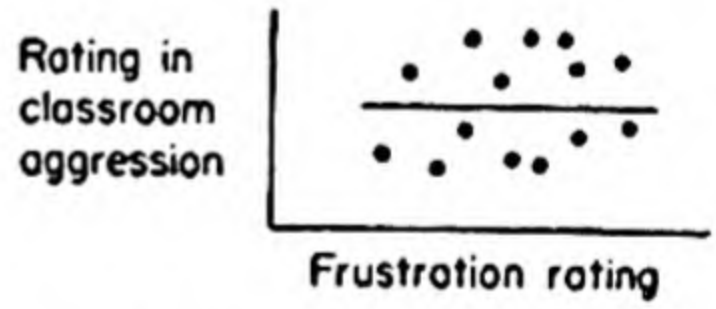


Fig. 8.9. The line of regression for the two variables is a horizontal line through the mean ($r = 0$).

A more workable computing formula for the coefficient of correlation than the formula given above is the Pearsonian product-moment coefficient of correlation, defined as:

$$r = \frac{\Sigma xy}{n_x \sigma_y} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} \tag{30}$$

$$= \frac{19}{\sqrt{(30)(18)}}$$

$$= \frac{19}{\sqrt{540}}$$

$$= \frac{19}{23.2} \doteq .82$$

where $x = X - \overline{X}$, and $y = Y - \overline{Y}$. This formula for the coefficient of correlation will yield the same results as the previous one if the relationship between the variables is assumed to be linear.

Note that the formula is symmetric. If X's are replaced by Y's, the formula is unchanged. Unlike regression, it does not matter in correlation which variable is called X or which variable Y. One variable is not considered the independent variable or the other, the dependent variable. If we regarded frustration rating as being dependent upon rating in classroom aggression (and not vice versa), although we would get a different line of regression, the coefficient of correlation would still be equal to .82.

It would be legitimate, in working out a *regression* problem, to predetermine what kind of distribution of frustration ratings (the independent variable) we want in our sample. Then, children would be selected at random within the allotted groups, and their aggression rating (the dependent variable) determined. In *correlation*, on the other hand, there can be no such predetermination unless the stratified proportions in each X-interval are representative of the universe.

The product moment correlation coefficient can be interpreted in terms of standard scores. The standard score for the X variable is x/σ_x, and for the Y variable it is y/σ_y. The arithmetic mean of the paired standard scores is r.

$$r = \frac{\Sigma}{n}\left(\frac{x}{\sigma_x}\right)\left(\frac{y}{\sigma_y}\right)$$

If each child had the same standard score in the frustration rating that he had in the rating in classroom aggression, there would be perfect correlation, equal to 1.

The correlation coefficient, like the mean and the proportion, is a sample statistic. To determine the probability of getting a correlation coefficient of .82 in a universe with *no correlation*, we would have to know the sampling distribution of the sample r's around a universe r of zero. When the sample size is greater than 30, the sampling dis-

tribution of r's approximates a normal distribution whose standard deviation is given by the formula:

$$\sigma_r = \frac{1}{\sqrt{n-1}} \tag{31}$$

Assume that our r of .82 came from a sample of 103 children. To determine the probability of getting a correlation of .82 when there is no correlation in the universe, we compute a standard score:

$$\text{standard score} = \frac{r_s - r_u}{\sigma_r} = \frac{.82 - 0}{\frac{1}{\sqrt{n-1}}} = \frac{.82}{\frac{1}{\sqrt{102}}} = \frac{.82}{\frac{1}{10.1}}$$

$$= 8.3 \text{ standard error units}$$

The standard score of 8.3 standard error units tells us that if there is no correlation in the universe of nursery school children between frustration rating and aggression rating, then the probability of getting a sample r of .82 is exceedingly small. We would consequently reject an hypothesis of no correlation.

We might next want to determine whether our sample correlation of .82 could have come from a universe with a correlation of .40. The sampling distribution for a universe correlation of .40 is skewed. (This is not unexpected, since the range of r extends from -1 to $+1$). R. A. Fisher has devised a formula for transforming the r into a Z-value that has an approximately normal sampling distribution, which is almost entirely independent of the universe value of r. The Z-transformation can be used for large or small hypothetical r's and for large or small samples. The standard error of Z is:

$$\sigma_z = \frac{1}{\sqrt{n-3}} \tag{32}$$

The table on page 176 gives the transformation of r to Z. The Z-value for an r of .82 is 1.157, for an r of .40 is .424. The standard error of the Z-transformation is equal to $1/\sqrt{n-3} = 1/\sqrt{100} = \frac{1}{10}$ = .10. Consequently, the standard score is 7.3 standard error units:

$$\frac{Z_s - Z_u}{\sigma_z} = \frac{1.157 - .424}{.10} = \frac{.733}{.10} = 7.3 \text{ standard error units}$$

The probability of getting an r of .82, due to random sampling fluctuations with a universe r of .40, is again exceedingly small, and so we reject an hypothetical universe value of .40.

We may next want to test the *significance of the difference* between our sample r of .82 from 103 nursery school children and a sample r of .60 from 153 junior high school students. We transform .82 into a Z-value of 1.157 and .60 into a Z-value of .693. The standard error of the difference between two Z-values is .13:

$$\sigma_{Z_1-Z_2} = \sqrt{\sigma_{Z_1}^2 + \sigma_{Z_2}^2} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}} = \sqrt{\frac{1}{100} + \frac{1}{150}}$$
$$= \sqrt{.0100 + .0067} = \sqrt{.0167}$$
$$= .13$$

The standard score is equal to more than 3 standard error units:

$$\frac{Z_1 - Z_2}{\sigma_{Z_1-Z_2}} = \frac{1.157 - .693}{.13} = \frac{.464}{.13} = 3.5 \text{ standard error units}$$

The difference in the frustration-aggression correlation between nursery school and junior high school children could have occurred less than 5 times in 10,000 through errors of random sampling if there were no difference between correlations in the two universes. Consequently, we reject the hypothesis of no difference in frustration-aggression correlation between nursery school and junior high school children.

There are some short-cut methods of solving the regression and correlation formulas, for use with calculators, in which large X's and Y's can be substituted for the small x and y deviations from the mean:

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2}$$

$$r = \frac{\Sigma xy}{n\sigma_x\sigma_y} = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{n\Sigma(X^2) - (\Sigma X)^2}\sqrt{n\Sigma(Y^2) - (\Sigma Y)^2}}$$

$$\sigma_{y_c} = \sqrt{\frac{\Sigma(Y - Y_c)^2}{n}} = \sigma_y\sqrt{1 - r^2}$$

Computation of Correlation Coefficient for Grouped Data. Two different interviewers have classified respondents by age. The age distribution given by the first interviewer is recorded in the margin of the vertical scale of Table 8-8 (f_y), the age distribution given by the second interviewer is classified in the margin of the horizontal scale (f_x). Each cell of the table contains the number of respondents clas-

sified simultaneously by the first and second interviewer within a certain age interval. There are, for example, 11 persons classified in age 30–39 by the first interviewer but in age 40–49 by the second interviewer. If for each cell frequency there is substituted an equivalent number of dots, we would have a scatter diagram of the frequency distribution. We might want to plot such a diagram to determine whether our data are linear.

The correlation coefficient would give some indication of the consistency of the respondent in stating his age, as well as the consistency of the observers in classifying the same respondents on age. The procedure used to compute a correlation coefficient for grouped data is as follows:

I. The row f_y gives the frequency of respondents for each age interval gotten by the first interviewer; the column f_x gives the same information for the second interviewer. The row is summed and the column is summed. Each is equal to total frequency ($\Sigma f_y = \Sigma f_x = n$).

II. The row and column y' and x' give the interval deviations from an arbitrary origin. The midpoint for the interval under 20 is assumed to be 15.

III. The row and column $f_y y'$ and $f_x x'$ are equal to I times II. The row and column III are summed.

IV. The row and column $f_y(y'^2)$ and $f_x(x'^2)$ are equal to II times III. The row and column IV are summed.

V. One term remains to be computed: $\Sigma f x' y'$. This is obtained by multiplying the frequency in each cell by the product of the cell's interval deviations. The results for each cell are encircled in the lower right hand corner of the cell and are then summed. A formula for the coefficient of correlation for grouped data:

$$r = \frac{n\Sigma f x' y' - (\Sigma f x')(\Sigma f y')}{\sqrt{n\Sigma f(x'^2) - (\Sigma f x')^2}\sqrt{n\Sigma f(y'^2) - (\Sigma f y')^2}} \tag{33}$$

Questions to Ask about a Coefficient of Correlation

1. Have we assumed a causal relationship when none exists?

The coefficient of correlation implies nothing about causation. We may find that over a period of time there has been increased financial aid to underdeveloped countries and also an increase in comedy-act television shows. The correlation may be almost perfect: as aid to underdeveloped countries increases, the proportion of comedy acts

Table 8-8. Coefficient of Correlation for Grouped Data

		*Classification by Second Interviewer** (*X*)								(I) *Total*	(II)	(III)	(IV)	(V)
		Under 20	20–29	30–39	40–49	50–59	60–69	70–79	80–89	f_v	y'	$f_v y'$	$f_v(y'^2)$	$fx'y'$
Classification by First Interviewer (*Y*)	80–89							$1_{(6)}$	$2_{(18)}$	3	3	9	27	24
	70–79							$4_{(16)}$		4	2	8	16	16
	60–69					$9_{(0)}$	$22_{(22)}$	$2_{(4)}$		33	1	33	33	26
	50–59			$1_{(0)}$	$7_{(0)}$	$25_{(0)}$	$4_{(0)}$			37	0	0	0	0
	40–49	$1_{(4)}$		$14_{(28)}$	$46_{(46)}$	$6_{(0)}$				67	−1	−67	67	78
	30–39		$4_{(24)}$	$65_{(260)}$	$11_{(22)}$	$2_{(0)}$	$2_{(-4)}$			84	−2	−168	336	302
	20–29	$1_{(12)}$	$41_{(369)}$	$11_{(66)}$	$7_{(21)}$					60	−3	−180	540	468
(I) *Total:*	f_x	2	45	91	71	42	28	7	2	288		−365	1,019	914
(II)	x'	−4	−3	−2	−1	0	1	2	3					
(III)	$f_x x'$	−8	−135	−182	−71	0	28	14	6	−348				
(IV)	$f_x(x'^2)$	32	405	364	71	0	28	28	18	946				
(V)	$fx'y'$	16	393	354	89	0	18	26	18	914				

$$r = \frac{n\Sigma fx'y' - (\Sigma fx')(\Sigma fy')}{\sqrt{n\Sigma f(x'^2) - (\Sigma fx')^2}\sqrt{n\Sigma f(y'^2) - (\Sigma fy')^2}} = \frac{(288)(914) - (-348)(-365)}{\sqrt{(288)(946) - (-348)^2}\sqrt{(288)(1019) - (-365)^2}}$$

$$= \frac{263{,}232 - 127{,}020}{\sqrt{272{,}448 - 121{,}104}\sqrt{293{,}472 - 133{,}225}} = \frac{136{,}212}{\sqrt{(151{,}344)(160{,}247)}} = \frac{136{,}212}{\sqrt{24{,}252{,}421{,}968}}$$

$$= \frac{136{,}212}{155{,}732} = \underline{.875}$$

increases. Both have increased over time; both are in part effects of the same phenomena. But there is no cause-effect relationship between them.

A coefficient of correlation says nothing about content. There is nothing inherent in correlation technique that will tell us whether or not one variable causes another. To assign causation, we must know the subject matter.

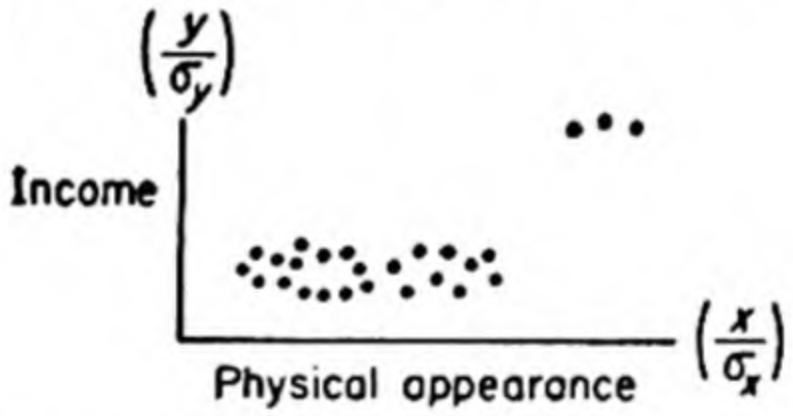
2. Does our correlation represent a relationship between overlapping data?

If we correlate grades of freshmen with grades of all students, we would expect to get a spuriously high correlation, since freshmen are included in the universe of all students.

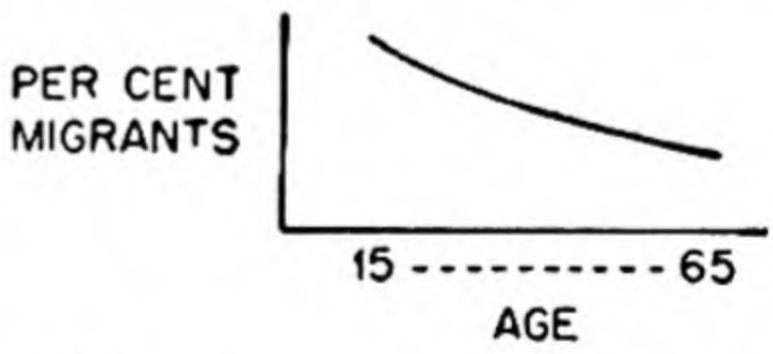
3. Is our high correlation caused by a few highly skewed cases?

One conception of a correlation is an arithmetic mean of the product of standard scores. A mean may be greatly influenced by a few highly skewed cases. Although there is actually no relationship, for example, between physical appearance and income (except possibly in Hollywood), the presence of a few highly skewed cases may result in a rather high correlation

$$r = \frac{\Sigma}{n}\left(\frac{x}{\sigma_x}\right)\left(\frac{y}{\sigma_y}\right)$$ (Mean of the product of standard scores)



4. Are we predicting to a universe to which our sample does not apply? Have we extrapolated the trend line and interpreted our data to apply to areas where it is inapplicable?



Suppose that we find a negative relation between age and per cent of migrants. The younger the age, the greater the per cent of migrants. Does this mean that among children under five we should find the greatest migrant percentage? Our data begins at the age 15–19, and to say something about the per cent of migrants under five would extend the range of our universe.

5. Are our data linear? A linear relationship is one in which a change of one unit in one variable always involves a certain constant change in the other variable.

Our definition of the coefficient of correlation r involves a regression line of the form $Y = a + bX$ from which squared vertical deviations are a minimum; r therefore measures the closeness of relationship between the two variables only when the trend of the data is approximately linear. If the relationship between two variables is nonlinear, even though it is high, the coefficient of correlation we have used might still approximate zero.

6. Is our correlation between group data or between personal data?

The correlation by city between number of libraries per capita and number of night clubs per capita may be high and positive. The correlation by person between library attendance and night club attendance will probably be substantially different.

The correlation by city reflects factors of urbanization; the correlation by person reflects personal habits. There is no reason why they must give the same results.

7. Could this sample correlation have been obtained purely by chance, because we have chosen one rather than another of the possible samples?

8. Have we controlled significant variables, which may crucially influence the results?

We may find, for example, a high correlation between intermarriage (i.e., marriage between individuals of different religions) and marital happiness. But perhaps those who intermarry regard themselves as very much in love; perhaps the marriage happiness test was given within a few years of marriage, before any children were born to the couples, so that the high correlation between intermarriage and happiness may be due not to the association between intermarriage and happiness but to the fact that couples very much in love, married only a few years (still in the honeymoon period of marriage) with no children, tend to regard themselves as happy, whether or not they have intermarried. If we control these factors, the relationship between intermarriage and marital happiness may almost vanish.

How do we find significant control variables? A scatter diagram can help in the search for significant relevant variables by revealing deviations from the trend. For example, in the study of the relation-

ship between the sex ratio of cities and the per cent of women married, the scatter diagram may look like this:

The scatter diagram reveals a high positive relationship—in cities where there is a high sex ratio there is also a high percentage of women married. But there are a few cities where this relationship does not hold. A study of these cities may reveal vast social status differences between the sexes, a distorted age-sex balance (e.g., the women may be old, the men young), or some other factor unique for these cities.

Partial correlation provides a method of controlling or holding constant significant variables conceptually.

9. How perfect are our methods of measurement? Are the rating scales of frustration and aggression, for example, reliable and valid?

Problems of reliability are concerned with the consistency with which an index or scale measures something. In a test of aggression, will each child, under similar conditions, get the same score regardless of administrator, time of day, and place? Will he get the same score in a test consisting of one-half the items, randomly selected and administered on successive days? Errors of measurement tend to reduce the correlation coefficient.

The problem of validity is a problem of definition. Are we really describing what we think we are describing? Is our definition consistent with some independent criterion purported to measure the same thing? In the aggression test, are we measuring the same characteristics in all children, or could the appearance of aggression in some children be merely the manifestation of a different social norm of behavior? If the results of this test are consistent with the determination of aggressiveness in nursery school children by a group of psychiatrists, the test may be considered validated.

Because (1) we have usually not controlled all relevant variables, and (2) our measurement of the variables we are using is rarely perfect, we shall rarely get a perfect correlation coefficient.

REFERENCES

Freund, John E., *Modern Elementary Statistics*, chap. 14. New York: Prentice-Hall, Inc., 1952.

Hagood, M. J., and D. O. Price, *Statistics for Sociologists*, chap. 23. New York: Henry Holt and Co., 1952.

McNemar, Quinn, *Psychological Statistics*, chaps. 6, 7 and 8. New York: John Wiley and Sons, Inc., 1949.

Mode, Elmer B., *The Elements of Statistics*, Rev. ed., chap. 11. New York: Prentice-Hall, Inc., 1951.

EXERCISES

1. One measure of the reliability of information on a socio-economic characteristic is the consistency with which different observers classify the same individuals on this characteristic. The table gives the economic-status classification of the same respondents by two different sets of interviewers.

Economic Status Classifications of the Same Respondents by Two Sets of Interviewers

		Classification by Second Interviewer					
		On Relief	*Poor*	*Average*	*Average-plus*	*Wealthy*	*Total*
Classification by First Interviewer	Wealthy	—	—	4	3	3	10
	Average-plus	—	—	21	20	4	45
	Average	—	5	74	24	2	105
	Poor	6	59	57	2	1	125
	On relief	9	9	8	—	—	26
	Total:	15	73	164	49	10	311

Source: Frederick Mosteller, "The Reliability of Interviewers Ratings," in *Gauging Public Opinion*, ed., Hadley Cantril (Princeton: Princeton University Press, 1947), p. 101.

(a) Compute the product moment correlation coefficient between the two economic classifications. Is economic status a discrete or continuous variable? What assumptions must be made in order to use a Pearsonian product-moment correlation coefficient for these data?

2. On what basis can the following correlations or lines of regression be criticized:

(a) The correlation between wages of school teachers and amount of liquor sold, 1900 to 1920, was found to be .96.

(b) To demonstrate that "blue-collar" workers' wages have kept pace with other salaries within a large manufacturing plant, a correlation is made between workers' annual take-home pay and total money disbursed annually in wages and salaries, from 1910 to 1950. The coefficient of correlation is found to be plus .80.

(c) Between 1932 and 1944 a consistently greater proportion of farmers voted for the Democratic party. It is therefore assumed that the farmer vote in 1948 should also be predominantly Democratic.

(d) The correlation between the score for neuroticism on a certain test

and the degree of neurosis in diagnosed neurotics is .60. This test is therefore used as a diagnostic tool in colleges.

(e) It is argued that divorce is bad since the correlation between juvenile delinquency and broken homes is plus .55.

3. When asked to give typical members of the upper and lower class, raters tend to give not the average but the extremes, those about whom there is no doubt as to class membership. How would this affect the coefficient of correlation between class membership and some other variables?

APPENDIX

Table		*Page*
	Formulas Appearing in the Text	162–165
I.	Squares to Four Significant Digits	166
II.	Area under the Normal Curve Corresponding to Distance on Baseline from Mean to Given Sigma-Distance	168
IIIa.	Distribution of t (Two-Tailed Probability)	169
IIIb.	Distribution of t (One-Tailed Probability)	170
IV.	Ninety-Five Per Cent Confidence Limits for Proportions	171
V.	Distribution of χ^2	172
VI.	Random Numbers	173
VII.	Values of Fisher's Z Corresponding to Values of r	175
VIII.	Ninety-Five Per Cent Confidence Limits for the Coefficient of Correlation	175

Formulas Appearing in the Text

Formula	*Description*	*Page*
$\frac{\Sigma X}{n}$	Arithmetic mean from ungrouped data	34
$\frac{\Sigma[(f)\cdot(\text{m.p.})]}{n}$	Arithmetic mean from grouped data	36
$\text{Guessed Mean} + \left(\frac{\Sigma fd'}{n}\right) C$	Arithmetic mean from grouped data (using guessed mean)	36
$\text{Lowest point in interval containing } \frac{1}{2}n\text{th case} + \left(\frac{\text{Frequency needed in median-interval to get to } \frac{1}{2}n}{\text{Frequency in median interval}}\right) \times \text{Size of interval}$	Median from grouped data	38
Highest value minus lowest value	Range	44
$\frac{\Sigma\lvert x\rvert}{n}$	Mean deviation	44
$\frac{\Sigma(x^2)}{n}$	Variance from ungrouped data	44
$\sqrt{\frac{\Sigma(x^2)}{n}}$	Standard deviation—ungrouped data (using deviations from the mean)	44
$\sqrt{\frac{\Sigma(X^2)}{n} - \left(\frac{\Sigma X}{n}\right)^2}$ or $\frac{1}{n}\sqrt{n\Sigma X^2 - (\Sigma X)^2}$	Standard deviation—ungrouped data (using observational values)	46

Formulas Appearing in the Text (*Cont'd*)

Formula	*Description*	*Page*
$C\sqrt{\frac{\Sigma f(d'^2)}{n} - \left(\frac{\Sigma fd'}{n}\right)^2}$	Standard deviation—grouped data (using guessed mean)	47
$\sqrt{\frac{\Sigma f(X^2)}{n} - \left(\frac{\Sigma fX}{n}\right)^2}$ or $\frac{1}{n}\sqrt{n\Sigma f(X^2) - (\Sigma fX)^2}$	Standard deviation—grouped data (using observational values)	47
$\frac{Q_3 - Q_1}{2}$	Semi-interquartile range	48
$\frac{X - \bar{X}}{s} = \frac{x}{s}$	Standard score	58
$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$	Standard error of the mean	87
est $\sigma_{\bar{X}} = \frac{s}{\sqrt{n-1}}$	Estimted standard error of the mean (using standard deviation of the sample)	97
$z = \frac{\bar{X} - M}{\sigma_{\bar{X}}}$	A z-score test of significance	97
est $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}$	Estimated standard error of the difference between two means	104
est $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{(s_p^2)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$	Estimated standard error of the difference between two means	104

Formulas Appearing in the Text (*Cont'd*)

Formula	*Description*	*Page*
$\frac{(\bar{X} - M)}{s}\sqrt{n-1}$	..... *t*-Test	106
$\sigma_{p_s} = \sqrt{\frac{p_u q_u}{n}}$	..... Standard error of a proportion	110
$\sigma_{p_{s_1}} - p_{s_2} = \sqrt{\frac{p_{u_1} q_{u_1}}{n_1} + \frac{p_{u_2} q_{u_2}}{n_2}}$	..... Standard error of the difference between two proportions	118
$\sigma_{p_{s_1}} - p_{s_2} = \sqrt{p_u q_u \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$	..... Standard error of the difference between two proportions	118
$p_u = \frac{n_1 p_{s_1} + n_2 p_{s_2}}{n_1 + n_2}$	..... Estimate of universe proportion	118
$\chi^2 = \Sigma \frac{(f_0 - f_e)^2}{f_e}$	..... Chi-square	125
$T^2 = \frac{\chi^2}{n\sqrt{(s-1)(t-1)}}$	..... A measure of degree of association	134
$b = \frac{\Sigma xy}{\Sigma x^2}$	..... Slope for line of best fit	143
$b = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}}{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}$ or $\frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2}$	... Slope for line of best fit (using large X's and Y's)	144

Formulas Appearing in the Text (*Cont'd*)

Formula	*Description*	*Page*
$a = \bar{Y} - b\bar{X}$	Y-intercept for line of best fit	143
$s_{y \cdot x} = \sqrt{\frac{\Sigma(Y - Y_c)^2}{n}}$	Standard error of estimate of sample	145
$\text{est } \sigma_{y \cdot x} = \sqrt{\frac{\Sigma(Y - Y_c)^2}{n - 2}}$	Estimate of standard error of estimate for universe	146
$r^2 = 1 - \frac{\sigma_{y_c}^2}{\sigma_y^2}$	Coefficient of determination	148
$r = \sqrt{1 - \frac{\sigma_{y_c}^2}{\sigma_y^2}}$	Coefficient of correlation	148
$r = \frac{\Sigma xy}{n\sigma_x\sigma_y}$	Pearsonian product moment coefficient of correlation	148
$r = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{n\Sigma(X^2) - (\Sigma X)^2}\sqrt{n\Sigma(Y^2) - (\Sigma Y)^2}}$	Coefficient of linear correlation (using large X's and Y's)	152
$r = \frac{n\Sigma fx'y' - (\Sigma fx')(\Sigma fy')}{\sqrt{n\Sigma f(x'^2) - (\Sigma fx')^2}\sqrt{n\Sigma f(y'^2) - (\Sigma fy')^2}}$	Linear correlation coefficient for grouped data	153
$\sigma_r = \frac{1}{\sqrt{n - 1}}$	Standard error of r ($r_u = 0$)	151
$\sigma_z = \frac{1}{\sqrt{n - 3}}$	Standard error of Z	151
$\sigma_{z_1 - z_2} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$	Standard error of the difference between two Z-values	152

Table I

Squares to Four Significant Digits

(Square roots may be found by inverse interpolation)

N	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.0	1.000	1.020	1.040	1.061	1.082	1.102	1.124	1.145	1.166	1.188
1.1	1.210	1.232	1.254	1.277	1.300	1.322	1.346	1.369	1.392	1.416
1.2	1.440	1.464	1.488	1.513	1.538	1.562	1.588	1.613	1.638	1.664
1.3	1.690	1.716	1.742	1.769	1.796	1.822	1.850	1.877	1.904	1.932
1.4	1.960	1.988	2.016	2.045	2.074	2.102	2.132	2.161	2.190	2.220
1.5	2.250	2.280	2.310	2.341	2.372	2.402	2.434	2.465	2.496	2.528
1.6	2.560	2.592	2.624	2.657	2.690	2.722	2.756	2.789	2.822	2.856
1.7	2.890	2.924	2.958	2.993	3.028	3.062	3.098	3.133	3.168	3.204
1.8	3.240	3.276	3.312	3.349	3.386	3.422	3.460	3.497	3.534	3.572
1.9	3.610	3.648	3.686	3.725	3.764	3.802	3.842	3.881	3.920	3.960
2.0	4.000	4.040	4.080	4.121	4.162	4.202	4.244	4.285	4.326	4.368
2.1	4.410	4.452	4.494	4.537	4.580	4.622	4.666	4.709	4.752	4.796
2.2	4.840	4.884	4.928	4.973	5.018	5.062	5.108	5.153	5.198	5.244
2.3	5.290	5.336	5.382	5.429	5.476	5.522	5.570	5.617	5.664	5.712
2.4	5.760	5.808	5.856	5.905	5.954	6.002	6.052	6.101	6.150	6.200
2.5	6.250	6.300	6.350	6.401	6.452	6.502	6.554	6.605	6.656	6.708
2.6	6.760	6.812	6.864	6.917	6.970	7.022	7.076	7.129	7.182	7.236
2.7	7.290	7.344	7.398	7.453	7.508	7.562	7.618	7.673	7.728	7.784
2.8	7.840	7.896	7.952	8.009	8.066	8.122	8.180	8.237	8.294	8.352
2.9	8.410	8.468	8.526	8.585	8.644	8.702	8.762	8.821	8.880	8.940
3.0	9.000	9.060	9.120	9.181	9.242	9.302	9.364	9.425	9.486	9.548
3.1	9.610	9.672	9.734	9.797	9.860	9.922	9.986	10.05	10.11	10.18
3.2	10.24	10.30	10.37	10.43	10.50	10.56	10.63	10.69	10.76	10.82
3.3	10.89	10.96	11.02	11.09	11.16	11.22	11.29	11.36	11.42	11.49
3.4	11.56	11.63	11.70	11.76	11.83	11.90	11.97	12.04	12.11	12.18
3.5	12.25	12.32	12.39	12.46	12.53	12.60	12.67	12.74	12.82	12.89
3.6	12.96	13.03	13.10	13.18	13.25	13.32	13.40	13.47	13.54	13.62
3.7	13.69	13.76	13.84	13.91	13.99	14.06	14.14	14.21	14.29	14.36
3.8	14.44	14.52	14.59	14.67	14.75	14.82	14.90	14.98	15.05	15.13
3.9	15.21	15.29	15.37	15.44	15.52	15.60	15.68	15.76	15.84	15.92
4.0	16.00	16.08	16.16	16.24	16.32	16.40	16.48	16.56	16.65	16.73
4.1	16.81	16.89	16.97	17.06	17.14	17.22	17.31	17.39	17.47	17.56
4.2	17.64	17.72	17.81	17.89	17.98	18.06	18.15	18.23	18.32	18.40
4.3	18.49	18.58	18.66	18.75	18.84	18.92	19.01	19.10	19.18	19.27
4.4	19.36	19.45	19.54	19.62	19.71	19.80	19.89	19.98	20.07	20.16
4.5	20.25	20.34	20.43	20.52	20.61	20.70	20.79	20.88	20.98	21.07
4.6	21.16	21.25	21.34	21.44	21.53	21.62	21.72	21.81	21.90	22.00
4.7	22.09	22.18	22.28	22.37	22.47	22.56	22.66	22.75	22.85	22.94
4.8	23.04	23.14	23.23	23.33	23.43	23.52	23.62	23.72	23.81	23.91
4.9	24.01	24.11	24.21	24.30	24.40	24.50	24.60	24.70	24.80	24.90
5.0	25.00	25.10	25.20	25.30	25.40	25.50	25.60	25.70	25.81	25.91
N	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09

Table I (*Cont'd*)

N	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
5.0	25.00	25.10	25.20	25.30	25.40	25.50	25.60	25.70	25.81	25.91
5.1	26.01	26.11	26.21	26.32	26.42	26.52	26.63	26.73	26.83	26.94
5.2	27.04	27.14	27.25	27.35	27.46	27.56	27.67	27.77	27.88	27.98
5.3	28.09	28.20	28.30	28.41	28.52	28.62	28.73	28.84	28.94	29.05
5.4	29.16	29.27	29.38	29.48	29.59	29.70	29.81	29.92	30.03	30.14
5.5	30.25	30.36	30.47	30.58	30.69	30.80	30.91	31.02	31.14	31.25
5.6	31.36	31.47	31.58	31.70	31.81	31.92	32.04	32.15	32.26	32.38
5.7	32.49	32.60	32.72	32.83	32.95	33.06	33.18	33.29	33.41	33.52
5.8	33 64	33.76	33.87	33.99	34.11	34.22	34.34	34.46	34.57	34.69
5.9	34.81	34.93	35.05	35.16	35.28	35.40	35.52	35.64	35.76	35.88
6.0	36.00	36.12	36.24	36.36	36.48	36.60	36.72	36.84	36.97	37.09
6.1	37.21	37.33	37 45	37.58	37.70	37.82	37.95	38.07	38.19	38.32
6.2	38.44	38.56	38.69	38.81	38.94	39.06	39.19	39.31	39.44	39.56
6.3	39.69	39.82	39.94	40.07	40.20	40.32	40.45	40.58	40.70	40.83
6.4	40.96	41.09	41.22	41.34	41.47	41 60	41.73	41.86	41.99	42.12
6.5	42.25	42.38	42.51	42.64	42.77	42.90	43.03	43.16	43.30	43.43
6.6	43.56	43.69	43.82	43.96	44.09	44.22	44.36	44.49	44.62	44.76
6.7	44.89	45.02	45.16	45.29	45.43	45.56	45.70	45.83	45.97	46.10
6.8	46.24	46.38	46.51	46.65	46.79	46.92	47.06	47.20	47.33	47.47
6.9	47.61	47.75	47.89	48.02	48.16	48.30	48.44	48.58	48.72	48.86
7.0	49.00	49.14	49.28	49.42	49.56	49.70	49.84	49.98	50.13	50.27
7.1	50.41	50.55	50.69	50.84	50.98	51.12	51.27	51.41	51.55	51.70
7.2	51.84	51.98	52.13	52.27	52.42	52.56	52.71	52.85	53.00	53.14
7.3	53.29	53.44	53.58	53.73	53.88	54.02	54.17	54.32	54.46	54.61
7.4	54.76	54.91	55.06	55.20	55.35	55.50	55.65	55.80	55.95	56.10
7.5	56.25	56.40	56.55	56.70	56.85	57.00	57.15	57.30	57.46	57.61
7.6	57.76	57.91	58.06	58.22	58.37	58.52	58.68	58.83	58.98	59.14
7.7	59.29	59.44	59.60	59.75	59.91	60.06	60.22	60.37	60.53	60.68
7.8	60.84	61.00	61.15	61.31	61.47	61.62	61.78	61.94	62.09	62.25
7.9	62.41	62.57	62.73	62.88	63.04	63.20	63.36	63.52	63.68	63.84
8.0	64.00	64.16	64.32	64.48	64.64	64.80	64.96	65.12	65.29	65.45
8.1	65.61	65.77	65.93	66.10	66.26	66.42	66.59	66.75	66.91	67.08
8.2	67.24	67.40	67.57	67.73	67.90	68.06	68.23	68.39	68.56	68.72
8.3	68.89	69.06	69.22	69.39	69.56	69.72	69.89	70.06	70.22	70.39
8.4	70.56	70.73	70.90	71.06	71.23	71.40	71.57	71.74	71.91	72.08
8.5	72.25	72.42	72.59	72.76	72.93	73.10	73.27	73.44	73.62	73.79
8.6	73.96	74.13	74.30	74.48	74.65	74.82	75.00	75.17	75.34	75.52
8.7	75.69	75.86	76.04	76.21	76.39	76.56	76.74	76.91	77.09	77.26
8.8	77.44	77.62	77.79	77.97	78.15	78.32	78.50	78.68	78.85	79.03
8.9	79.21	79.39	79.57	79.74	79.92	80.10	80.28	80.46	80.64	80.82
9.0	81.00	81.18	81.36	81.54	81.72	81.90	82.08	82.26	82.45	82.63
9.1	82.81	82.99	83.17	83.36	83.54	83.72	83.91	84.09	84.27	84.46
9.2	84.64	84.82	85.01	85.19	85.38	85.56	85.75	85.93	86.12	86.30
9.3	86.49	86.68	86.86	87.05	87.24	87.42	87.61	87.80	87 98	88.17
9.4	88.36	88.55	88.74	88.92	89.11	89 30	89.49	89.68	89.87	90.06
9.5	90.25	90.44	90.63	90.82	91.01	91.20	91.39	91.58	91.78	91.97
9.6	92.16	92.35	92.54	92.74	92.93	93.12	93.32	93.51	93.70	93.90
9.7	94.09	94.28	94.48	94.67	94.87	95.06	95.26	95.45	95.65	95.84
9.8	96.04	96.24	96.43	96.63	96.83	97.02	97.22	97.42	97 61	97.81
9.9	98.01	98.21	98.41	98.60	98.80	99.00	99.20	99.40	99.60	99.80
N	**.00**	**.01**	**.02**	**.03**	**.04**	**.05**	**.06**	**.07**	**.08**	**.09**

SOURCE: Elmer Mode, *Elements of Statistics*, 2nd ed. (New York: Prentice-Hall, Inc., 1951), Table B, p. 364.

Table II

Area under the Normal Curve Corresponding to Distance on Baseline from Mean to Given Sigma-Distance

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0·0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	4989	.4989	.4989	.4990	.4990

SOURCE: J. Neyman, *First Course in Probability and Statistics* (New York: Henry Holt and Company, Inc. 1950).

Table IIIa

Distribution of *t*. The Probability of Exceeding Given Values of *t* by Number of Degrees of Freedom

(Two-tailed probability)

Degrees of Freedom (df)	Probability												
	·9	·8	·7	·6	·5	·4	·3	·2	·1	·05	·02	·01	001
1	·158	·325	·510	·727	1·000	1·376	1 963	3·078	6·314	12·706	31·821	63·657	636·619
2	·142	·289	·445	·617	·816	1·061	1·386	1·886	2·920	4·303	6·965	9·925	31.598
3	·137	·277	·424	·584	·765	·978	1·250	1·638	2·353	3·182	4·541	5·841	12.941
4	·134	271	·414	·569	·741	·941	1·190	1·533	2·132	2·776	3·747	4.604	8·610
5	·132	·267	·408	·559	·727	·920	1·156	1·476	2·015	2·571	3·365	4·032	6·859
6	·131	·265	·404	·553	·718	·906	1·134	1·440	1·943	2·447	3·143	3·707	5·959
7	·130	·263	·402	·549	·711	·896	1·119	1·415	1·895	2·365	2·998	3·499	5·405
8	·130	·262	·399	·546	·706	·889	1·108	1·397	1·860	2·306	2·896	3·355	5·041
9	·129	·261	·398	·543	·703	·883	1·100	1·383	1·833	2·262	2·821	3·250	4·781
10	·129	·260	·397	·542	·700	·879	1·093	1·372	1·812	2·228	2·764	3·169	4·587
11	·129	·260	·396	·540	·697	·876	1·088	1 363	1·796	2 201	2·718	3·106	4·437
12	·128	·259	·395	·539	·695	·873	1·083	1·356	1·782	2·179	2·681	3·055	4·318
13	·128	·259	·394	·538	·694	·870	1 079	1·350	1·771	2·160	2·650	3·012	4·221
14	·128	·258	·393	·537	·692	·868	1·076	1·345	1·761	2·145	2·624	2·977	4·140
15	·128	·258	·393	·536	·691	866	1·074	1·341	1·753	2·131	2·602	2·947	4·073
16	·128	·258	·392	·535	.690	·865	1·071	1·337	1·746	2·120	2·583	2·921	4·015
17	·128	·257	·392	·534	·689	·863	1·069	1·333	1·740	2·110	2·567	2·898	3·965
18	·127	·257	·392	·534	·688	·862	1·067	1·330	1·734	2·101	2·552	2·878	3·922
19	·127	·257	·391	·533	·688	·861	1·066	1·328	1·729	2·093	2·539	2·861	3·883
20	·127	·257	·391	·533	·687	·860	1·064	1·325	1·725	2·086	2·528	2·845	3·850
21	·127	·257	·391	·532	·686	·859	1·063	1·323	1·721	2·080	2·518	2·831	3·819
22	·127	·256	·390	·532	·686	·858	1·061	1·321	1·717	2·074	2·508	2·819	3·792
23	·127	·256	·390	·532	·685	·858	1·060	1·319	1·714	2·069	2·500	2·807	3·767
24	·127	256	·390	·531	·685	·857	1·059	1·318	1·711	2·064	2·492	2·797	3·745
25	·127	·256	·390	·531	·684	·856	1·058	1·316	1·708	2·060	2·485	2·787	3·725
26	·127	·256	390	·531	·684	·856	1·058	1·315	1·706	2·056	2·479	2·779	3·707
27	·127	·256	·389	·531	·684	·855	1·057	1·314	1·703	2·052	2·473	2·771	3·690
28	·127	·256	·389	·530	·683	·855	1·056	1·313	1·701	2·048	2·467	2·763	3·674
29	·127	·256	·389	·530	·683	·854	1·055	1·311	1·699	2·045	2·462	2·756	3·659
30	·127	·256	·389	·530	·683	·854	1·055	1·310	1·697	2·042	2·457	2·750	3·646
40	·126	·255	·388	·529	·681	·851	1·050	1·303	1·684	2·021	2·423	2·704	3·551
60	·126	·254	·387	·527	·679	·848	1·046	1·296	1·671	2·000	2·390	2·660	3·460
120	·126	·254	·386	·526	·677	·845	1·041	1·289	1·658	1·980	2·358	2·617	3·373
∞	·126	·253	·385	·524	·674	·842	1·036	1·282	1·645	1·960	2·326	2·576	3·291

SOURCE: Reprinted from Table III of Fisher and Yates: *Statistical Tables for Biological, Agricultural, and Medical Research*, 4th ed., 1953, published by Oliver and Boyd Limited, Edinburgh, by permission of the authors and publishers.

Table IIIb

Distribution of *t*. The Probability of Exceeding Given Positive Values of *t* by Number of Degrees of Freedom

(One-tailed probability)

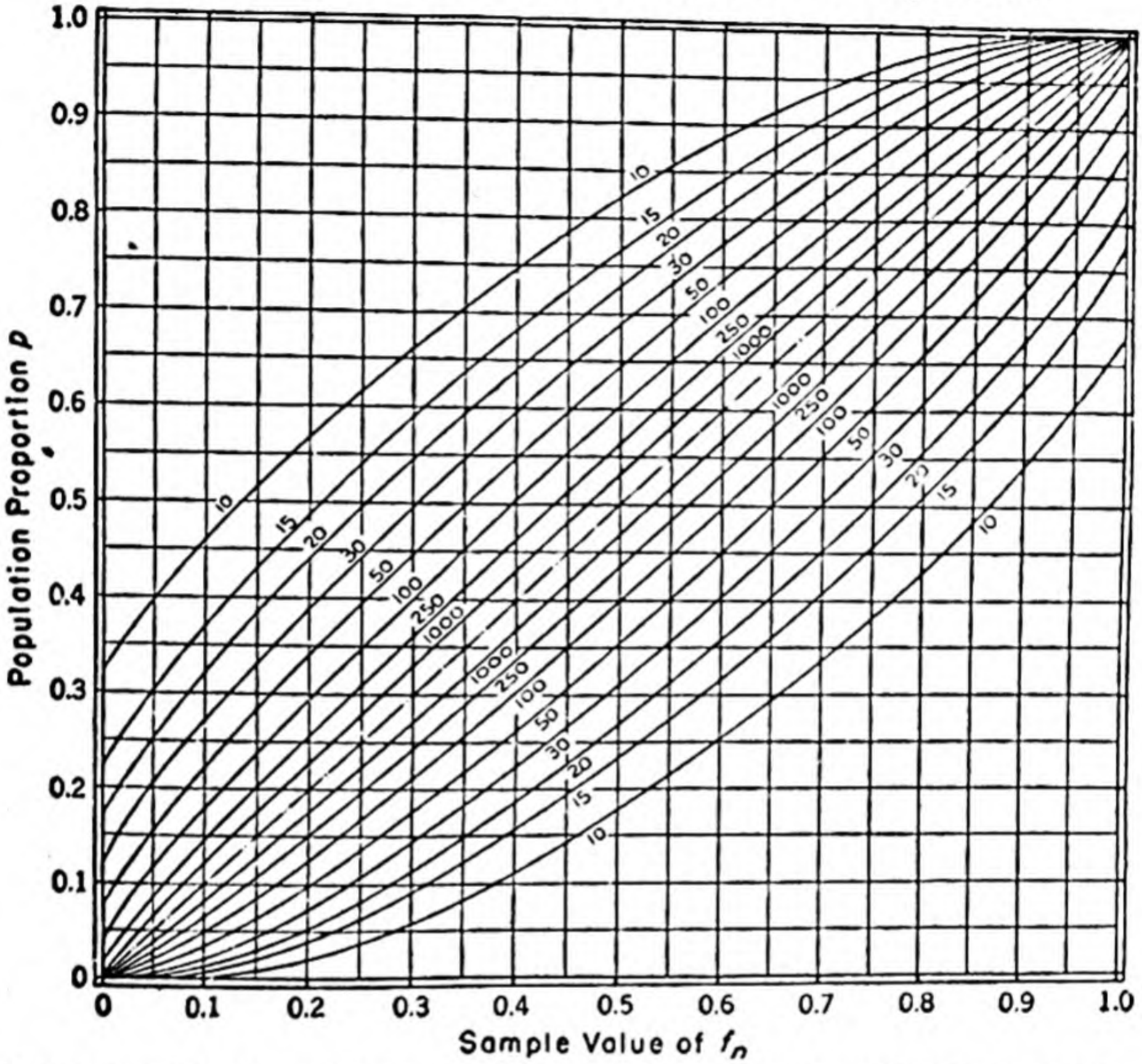
Probability

Degrees of Freedom (df)	.45	.40	.35	.20	.15	.10	.05	.025	.01	.005
1	.158	.325	.510	1.376	1.963	3.078	6 314	12.71	31 82	63.66
2	.142	289	.445	1.061	1.386	1 886	2.920	4.303	6.965	9.925
3	.137	.277	.424	.978	1.250	1.638	2.353	3.182	4.541	5.841
4	.134	.271	414	.941	1.190	1 533	2 132	2.776	3 747	4.604
5	.132	.267	408	.920	1.156	1.476	2.015	2.571	3.365	4.032
6	.131	.265	404	.906	1.134	1.440	1.943	2.447	3.143	3.707
7	.130	.263	.402	.896	1.119	1.415	1.895	2.365	2.998	3.499
8	.130	.262	.399	.889	1.108	1.397	1.860	2.306	2 896	3.355
9	.129	261	398	.883	1.100	1.383	1 833	2.262	2.821	3.250
10	.129	.260	397	.879	1.093	1.372	1.812	2.228	2.764	3.169
11	.129	.260	.396	.876	1.088	1.363	1 796	2.201	2.718	3.106
12	128	259	.395	.873	1.083	1.356	1.782	2.179	2.681	3.055
13	.128	.259	.394	.870	1.079	1.350	1.771	2.160	2.650	3.012
14	.128	.258	393	.868	1.076	1.345	1.761	2.145	2.624	2.977
15	.128	.258	393	.866	1.074	1 341	1.753	2.131	2.602	2.947
16	128	.258	.392	.865	1.071	1.337	1.746	2.120	2.583	2.921
17	.128	.257	.392	.863	1.069	1.333	1.740	2.110	2.567	2.898
18	.127	.257	.392	862	1.067	1.330	1.734	2.101	2.552	2.878
19	.127	.257	391	.861	1.066	1.328	1.729	2.093	2.539	2.861
20	.127	.257	.391	.860	1 064	1.325	1.725	2.086	2.528	2.845
21	.127	257	.391	.859	1.063	1.323	1.721	2.080	2.518	2 831
22	127	256	.390	.858	1.061	1.321	1.717	2.074	2.508	2.819
23	.127	.256	.390	.858	1.060	1.319	1.714	2.069	2.500	2.807
24	.127	.256	.390	.857	1.059	1.318	1.711	2.064	2.492	2.797
25	127	.256	.390	.856	1.058	1.316	1.708	2.060	2.485	2.787
26	127	.256	.390	.856	1.058	1.315	1.706	2.056	2.479	2.779
27	.127	.256	.389	.855	1.057	1.314	1.703	2.052	2.473	2.771
28	.127	.256	389	.855	1.056	1.313	1.701	2.048	2.467	2.763
29	127	.256	.389	.854	1.055	1.311	1.699	2.045	2 462	2.756
30	127	256	.389	.854	1.055	1.310	1 697	2.042	2.457	2.750
40	.126	.255	.388	.851	1.050	1.303	1 684	2.021	2.423	2.704
60	126	.254	.387	.848	1.046	1.296	1.671	2.000	2.390	2.660
120	126	.254	386	.845	1.041	1.289	1.658	1.980	2.358	2 617
∞	.126	.253	.385	.842	1.036	1.282	1.645	1.960	2.326	2.576
Degrees of Freedom (df)	.45	.40	.35	.20	.15	.10	.05	.025	.01	.005

SOURCE: Abridged from Table III of Fisher and Yates: *Statistical Tables for Biological, Agricultural, and Medical Research*, 4th ed., 1953, published by Oliver and Boyd Limited, Edinburgh, by permission of the authors and publishers.

Table IV

Ninety-Five Per Cent Confidence Limits for Proportions



SOURCE: Reprinted from C. J. Clopper and E. S. Pearson, "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," *Biometrika*, Vol. 26 (1934), by permission of Professor E. S. Pearson.

Table V

Distribution of χ^2

(The probability of exceeding given values of χ^2 by number of degrees of freedom)

Degrees of Freedom (df)	Probability													
	·99	·98	·95	·90	·80	·70	50	30	20	10	·05	·02	·01	·001
1	·0³157	·0³628	·00393	·0158	·0642	·148	·455	1·074	1·642	2·706	3·841	5·412	6·635	10·827
2	·0201	·0404	·103	·211	·446	·713	1·386	2.408	3·219	4·605	5 .991	7·824	9·210	13·815
3	·115	·185	·352	·584	1·005	1·424	2·366	3·665	4·642	6·251	7·815	9·837	11·341	16·268
4	·297	·429	·711	1·064	1·649	2 195	3.357	4·878	5·989	7·779	9·488	11·668	13·277	18·465
5	·554	·752	1·145	1·610	2·343	3·000	4·351	6·064	7·289	9·236	11·070	13·388	15·086	20·517
6	·872	1·134	1·635	2 204	3·070	3·828	5·348	7·231	8·558	10·645	12·592	15·033	16·812	22·457
7	1·239	1·564	2·167	2·833	3·822	4·671	6·346	8·383	9·803	12·017	14·067	16·622	18·475	24·322
8	1·646	2·032	2·733	3·490	4·594	5.527	7.344	9·524	11·030	13·362	15·507	18·168	20·090	26·125
9	2·088	2·532	3·325	4·168	5·380	6·393	8·343	10·656	12·242	14·684	16·919	19·679	21·666	27·877
10	2·558	3·059	3·940	4·865	6·179	7·267	9·342	11·781	13·442	15·987	18·307	21·161	23·209	29·588
11	3·053	3·609	4·575	5·578	6·989	8·148	10·341	12·899	14·631	17·275	19·675	22·618	24·725	31·264
12	3.571	4·178	5·226	6·304	7·807	9·034	11·340	14·011	15·812	18·549	21·026	24·054	26·217	32·909
13	4·107	4·765	5·892	7·042	8·634	9·926	12·340	15·119	16·985	19·812	22·362	25·472	27·688	34·528
14	4·660	5·368	6·571	7·790	9·467	10·821	13·339	16·222	18·151	21·064	23·685	26·873	29·141	36·123
15	5·229	5·985	7·261	8·547	10·307	11·721	14·339	17·322	19·311	22·307	24 996	28·259	30·578	37·697
16	5·812	6·614	7·962	9·312	11·152	12·624	15 338	18·418	20·465	23·542	26·296	29·633	32·000	39·252
17	6·408	7·255	8·672	10·085	12·002	13·531	16·338	19·511	21·615	24·769	27·587	30·995	33·409	40·790
18	7·015	7·906	9·390	10·865	12·857	14·440	17·338	20·601	22·760	25·989	28·869	32·346	34·805	42·312
19	7·633	8·567	10·117	11·651	13·716	15·352	18·338	21·689	23·900	27·204	30·144	33·687	36·191	43·820
20	8·260	9·237	10·851	12·443	14·578	16·266	19·337	22 775	25·038	28·412	31·410	35·020	37·566	45·315
21	8·897	9·915	11·591	13·240	15·445	17·182	20·337	23·858	26·171	29·615	32·671	36·343	38·932	46·797
22	9·542	10·600	12·338	14·041	16·314	18·101	21·337	24·939	27·301	30·813	33·924	37·659	40·289	48·268
23	10·196	11·293	13·091	14·848	17·187	19·021	22·337	26·018	28·429	32·007	35·172	38·968	41·638	49·728
24	10·856	11·992	13·848	15·659	18·062	19·943	23·337	27·096	29·553	33·196	36·415	40·270	42·980	51·179
25	11·524	12·697	14·611	16·473	18·940	20·867	24·337	28·172	30·675	34·382	37·652	41·566	44·314	52·620
26	12·198	13·409	15 379	17·292	19·820	21·792	25·336	29·246	31·795	35·563	38·885	42·856	45·642	54·052
27	12·879	14·125	16·151	18·114	20·703	22·719	26·336	30·319	32·912	36·741	40·113	44·140	46·963	55·476
28	13·565	14·847	16·928	18·939	21·588	23·647	27·336	31·391	34·027	37·916	41·337	45·419	48·278	56·893
29	14·256	15·574	17·708	19·768	22·475	24·577	28·336	32·461	35·139	39·087	42·557	46·693	49·588	58·302
30	14·953	16 306	18·493	20·599	23·364	25·508	29·336	33·530	36·250	40·256	43·773	47·962	50·892	59·703

SOURCE: Reprinted from Table IV of Fisher and Yates: *Statistical Tables for Biological, Agricultural, and Medical Research*, 4th ed., 1953, published by Oliver and Boyd Limited, Edinburgh, by permission of the authors and publishers.

For degrees of freedom greater than 30, the expression $\sqrt{2\chi^2} - \sqrt{2(df) - 1}$ may be used as a normal deviate with unit variance, remembering that the probability for χ^2 corresponds with that of a single tail of the normal curve.

Table VI

Random Numbers

03 47 43 73 86	36 96 47 36 61	46 98 63 71 62	33 26 16 80 45	60 11 14 10 95
97 74 24 67 62	42 81 14 57 20	42 53 32 37 32	27 07 36 07 51	24 51 79 89 73
16 76 62 27 66	56 50 26 71 07	32 90 79 78 53	13 55 38 58 59	88 97 54 14 10
12 56 85 99 26	96 96 68 27 31	05 03 72 93 15	57 12 10 14 21	88 26 49 81 76
55 59 56 35 64	38 54 82 46 22	31 62 43 09 90	06 18 44 32 53	23 83 01 30 30
16 22 77 94 39	49 54 43 54 82	17 37 93 23 78	87 35 20 96 43	84 26 34 91 64
84 42 17 53 31	57 24 55 06 88	77 04 74 47 67	21 76 33 50 25	83 92 12 06 76
63 01 63 78 59	16 95 55 67 19	98 10 50 71 75	12 86 73 58 07	44 39 52 38 79
33 21 12 34 29	78 64 56 07 82	52 42 07 44 38	15 51 00 13 42	99 66 02 79 54
57 60 86 32 44	09 47 27 96 54	49 17 46 09 62	90 52 84 77 27	08 02 73 43 28
18 18 07 92 46	44 17 16 58 09	79 83 86 19 62	06 76 50 03 10	55 23 64 05 05
26 62 38 97 75	84 16 07 44 99	83 11 46 32 24	20 14 85 88 45	10 93 72 88 71
23 42 40 64 74	82 97 77 77 81	07 45 32 14 08	32 98 94 07 72	93 85 79 10 75
52 36 28 19 95	50 92 26 11 97	00 56 76 31 38	80 22 02 53 53	86 60 42 04 53
37 85 94 35 12	83 39 50 08 30	42 34 07 96 88	54 42 06 87 98	35 85 29 48 39
70 29 17 12 13	40 33 20 38 26	13 89 51 03 74	17 76 37 13 04	07 74 21 19 30
56 62 18 37 35	96 83 50 87 75	97 12 25 93 47	70 33 24 03 54	97 77 46 44 80
99 49 57 22 77	88 42 95 45 72	16 64 36 16 00	04 43 18 66 79	94 77 24 21 90
16 08 15 04 72	33 27 14 34 09	45 59 34 68 49	12 72 07 34 45	99 27 72 95 14
31 16 93 32 43	50 27 89 87 19	20 15 37 00 49	52 85 66 60 44	38 68 88 11 80
68 34 30 13 70	55 74 30 77 40	44 22 78 84 26	04 33 46 09 52	68 07 97 06 57
74 57 25 65 76	59 29 97 68 60	71 91 38 67 54	13 58 18 24 76	15 54 55 95 52
27 42 37 86 53	48 55 90 65 72	96 57 69 36 10	96 46 92 42 45	97 60 49 04 91
00 39 68 29 61	66 37 32 20 30	77 84 57 03 29	10 45 65 04 26	11 04 96 67 24
29 94 98 94 24	68 49 69 10 82	53 75 91 93 30	34 25 20 57 27	40 48 73 51 92
16 90 82 66 59	83 62 64 11 12	67 19 00 71 74	60 47 21 29 68	02 02 37 03 31
11 27 94 75 06	06 09 19 74 66	02 94 37 34 02	76 70 90 30 86	38 45 94 30 38
35 24 10 16 20	33 32 51 26 38	79 78 45 04 91	16 92 53 56 16	02 75 50 95 98
38 23 16 86 38	42 38 97 01 50	87 75 66 81 41	40 01 74 91 62	48 51 84 08 32
31 96 25 91 47	96 44 33 49 13	34 86 82 53 91	00 52 43 48 85	27 55 26 89 62
66 67 40 67 14	64 05 71 95 86	11 05 65 09 68	76 83 20 37 90	57 16 00 11 66
14 90 84 45 11	75 73 88 05 90	52 27 41 14 86	22 98 12 22 08	07 52 74 95 80
68 05 51 18 00	33 96 02 75 19	07 60 62 93 55	59 33 82 43 90	49 37 38 44 59
20 46 78 73 90	97 51 40 14 02	04 02 33 31 08	39 54 16 49 36	47 95 93 13 30
64 19 58 97 79	15 06 15 93 20	01 90 10 75 06	40 78 78 89 62	02 67 74 17 33
05 26 93 70 60	22 35 85 15 13	92 03 51 59 77	59 56 78 06 83	52 91 05 70 74
07 97 10 88 23	09 98 42 99 64	61 71 62 99 15	06 51 29 16 93	58 05 77 09 51
68 71 86 85 85	54 87 66 47 54	73 32 08 11 12	44 95 92 63 16	29 56 24 29 48
26 99 61 65 53	58 37 78 80 70	42 10 50 67 42	32 17 55 85 74	94 44 67 16 94
14 65 52 68 75	87 59 36 22 41	26 78 63 06 55	13 08 27 01 50	15 29 39 39 43
17 53 77 58 71	71 41 61 50 72	12 41 94 96 26	44 95 27 36 99	02 96 74 30 83
90 26 59 21 19	23 52 23 33 12	96 93 02 18 39	07 02 18 36 07	25 99 32 70 23
41 23 52 55 99	31 04 49 69 96	10 47 48 45 88	13 41 43 89 20	97 17 14 49 17
60 20 50 81 69	31 99 73 68 68	35 81 33 03 76	24 30 12 48 60	18 99 10 72 34
91 25 38 05 90	94 58 28 41 36	45 37 59 03 09	90 35 57 29 12	82 62 54 65 60
34 50 57 74 37	98 80 33 00 91	09 77 93 19 82	74 94 80 04 04	45 07 31 66 49
85 22 04 39 43	73 81 53 94 79	33 62 46 86 28	08 31 54 46 31	53 94 13 38 47
09 79 13 77 48	73 82 97 22 21	05 03 27 24 83	72 89 44 05 60	35 80 39 94 88
88 75 80 18 14	22 95 75 42 49	39 32 82 22 49	02 48 07 70 37	16 04 61 67 87
90 96 23 70 00	39 00 03 06 90	55 85 78 38 36	94 37 30 69 32	90 89 00 76 33

Table VI (*Cont'd*)

53 74 23 99 67	61 32 28 69 84	94 62 67 86 24	98 33 41 19 95	47 53 53 38 09
63 38 06 86 54	99 00 65 26 94	02 82 90 23 07	79 62 67 80 60	75 91 12 81 19
35 30 58 21 46	06 72 17 10 94	25 21 31 75 96	49 28 24 00 49	55 65 79 78 07
63 43 36 82 69	65 51 18 37 88	61 38 44 12 45	32 92 85 88 65	54 34 81 85 35
98 25 37 55 26	01 91 82 81 46	74 71 12 94 97	24 02 71 37 07	03 92 18 66 75
02 63 21 17 69	71 50 80 89 56	38 15 70 11 48	43 40 45 86 98	00 83 26 91 03
64 55 22 21 82	48 22 28 06 00	61 54 13 43 91	82 78 12 23 29	06 66 24 12 27
85 07 26 13 89	01 10 07 82 04	59 63 69 36 03	69 11 15 83 80	13 29 54 19 28
58 54 16 24 15	51 54 44 82 00	62 61 65 04 69	38 18 65 18 97	85 72 13 49 21
34 85 27 84 87	61 48 64 56 26	90 18 48 13 26	37 70 15 42 57	65 65 80 39 07
03 92 18 27 46	57 99 16 96 56	30 33 72 85 22	84 64 38 56 98	99 01 30 98 64
62 95 30 27 59	37 75 41 66 48	86 97 80 61 45	23 53 04 01 63	45 76 08 64 27
08 45 93 15 22	60 21 75 46 91	98 77 27 85 42	28 88 61 08 84	69 62 03 42 73
07 08 55 18 40	45 44 75 13 90	24 94 96 61 02	57 55 66 83 15	73 42 37 11 61
01 85 89 95 66	51 10 19 34 88	15 84 97 19 75	12 76 39 43 78	64 63 91 08 25
72 84 71 14 35	19 11 58 49 26	50 11 17 17 76	86 31 57 20 18	95 60 78 46 75
88 78 28 16 84	13 52 53 94 53	75 45 69 30 96	73 89 65 70 31	99 17 43 48 76
45 17 75 65 57	28 40 19 72 12	25 12 74 75 67	60 40 60 81 19	24 62 01 61 16
96 76 28 12 54	22 01 11 94 25	71 96 16 16 88	68 64 36 74 45	19 59 50 88 92
43 31 67 72 30	24 02 94 08 63	38 32 36 66 02	59 36 38 25 39	48 03 45 15 22
50 44 66 44 21	65 06 58 05 62	68 15 54 35 02	42 35 48 96 32	14 52 41 52 48
22 66 22 15 86	26 63 75 41 99	58 42 36 72 24	58 37 52 18 51	03 37 18 39 11
96 24 40 14 51	23 22 30 88 57	95 67 47 29 83	94 69 40 06 07	18 16 36 78 86
31 73 91 61 19	60 20 72 93 48	98 57 07 23 69	65 95 39 69 58	56 80 30 19 44
78 60 73 99 84	43 89 94 36 45	56 69 47 07 41	90 22 91 07 12	78 35 34 08 72
84 37 90 61 56	70 10 23 98 05	85 11 34 76 60	76 48 45 34 60	01 64 18 39 96
36 67 10 08 23	98 93 35 08 86	99 29 76 29 81	33 34 91 58 93	63 14 52 32 52
07 28 59 07 48	89 64 58 89 75	83 85 62 27 89	30 14 78 56 27	86 63 59 80 02
10 15 83 87 60	79 24 31 66 56	21 48 24 06 93	91 98 94 05 49	01 47 59 38 00
55 19 68 97 65	03 73 52 16 56	00 53 55 90 27	33 42 29 38 87	22 13 88 83 34
53 81 29 13 39	35 01 20 71 34	62 33 74 82 14	53 73 19 09 03	56 54 29 56 93
51 86 32 68 92	33 98 74 66 99	40 14 71 94 58	45 94 19 38 81	14 44 99 81 07
35 91 70 29 13	80 03 54 07 27	96 94 78 32 66	50 95 52 74 33	13 80 55 62 54
37 71 67 95 13	20 02 44 95 94	64 85 04 05 72	01 32 90 76 14	53 89 74 60 41
93 66 13 83 27	92 79 64 64 72	28 54 96 53 84	48 14 52 98 94	56 07 93 89 30
02 96 08 45 65	13 05 00 41 84	93 07 54 72 59	21 45 57 09 77	19 48 56 27 44
49 83 43 48 35	82 88 33 69 96	72 36 04 19 76	47 45 15 18 60	82 11 08 95 97
84 60 71 62 46	40 80 81 30 37	34 39 23 05 38	25 15 35 71 30	88 12 57 21 77
18 17 30 88 71	44 91 14 88 47	89 23 30 63 15	56 34 20 47 89	99 82 93 24 98
79 69 10 61 78	71 32 76 95 62	87 00 22 58 40	92 54 01 75 25	43 11 71 99 31
75 93 36 57 83	56 20 14 82 11	74 21 97 90 65	96 42 68 63 86	74 54 13 26 94
38 30 92 29 03	06 28 81 39 38	62 25 06 84 63	61 29 08 93 67	04 32 92 08 09
51 29 50 10 34	31 57 75 95 80	51 97 02 74 77	76 15 48 49 44	18 55 63 77 09
21 31 38 86 24	37 79 81 53 74	73 24 16 10 33	52 83 90 94 76	70 47 14 54 36
29 01 23 87 88	58 02 39 37 67	42 10 14 20 92	16 55 23 42 45	54 96 09 11 06
95 33 95 22 00	18 74 72 00 18	38 79 58 69 32	81 76 80 26 92	82 80 84 25 39
90 84 60 79 80	24 36 59 87 38	82 07 53 89 35	96 35 23 79 18	05 98 90 07 35
46 40 62 98 82	54 97 20 56 95	15 74 80 08 32	16 46 70 50 80	67 72 16 42 79
20 31 89 03 43	38 46 82 68 72	32 14 82 99 70	80 60 47 18 97	63 49 30 21 30
71 59 73 05 50	08 22 23 71 77	91 01 93 20 49	82 96 59 26 94	66 39 67 98 60

Source: Reprinted from Table XXXIII, (I) and (II), of Fisher and Yates: *Statistical Tables for Biological, Agricultural, and Medical Research*, 4th ed., 1953, published by Oliver and Boyd Limited, Edinburgh, by permission of the authors and publishers.

Table VII

Values of Fisher's Z Corresponding to Values of r

$$\log_e Z = \tfrac{1}{2} \log_e \frac{1 + r}{1 - r}$$

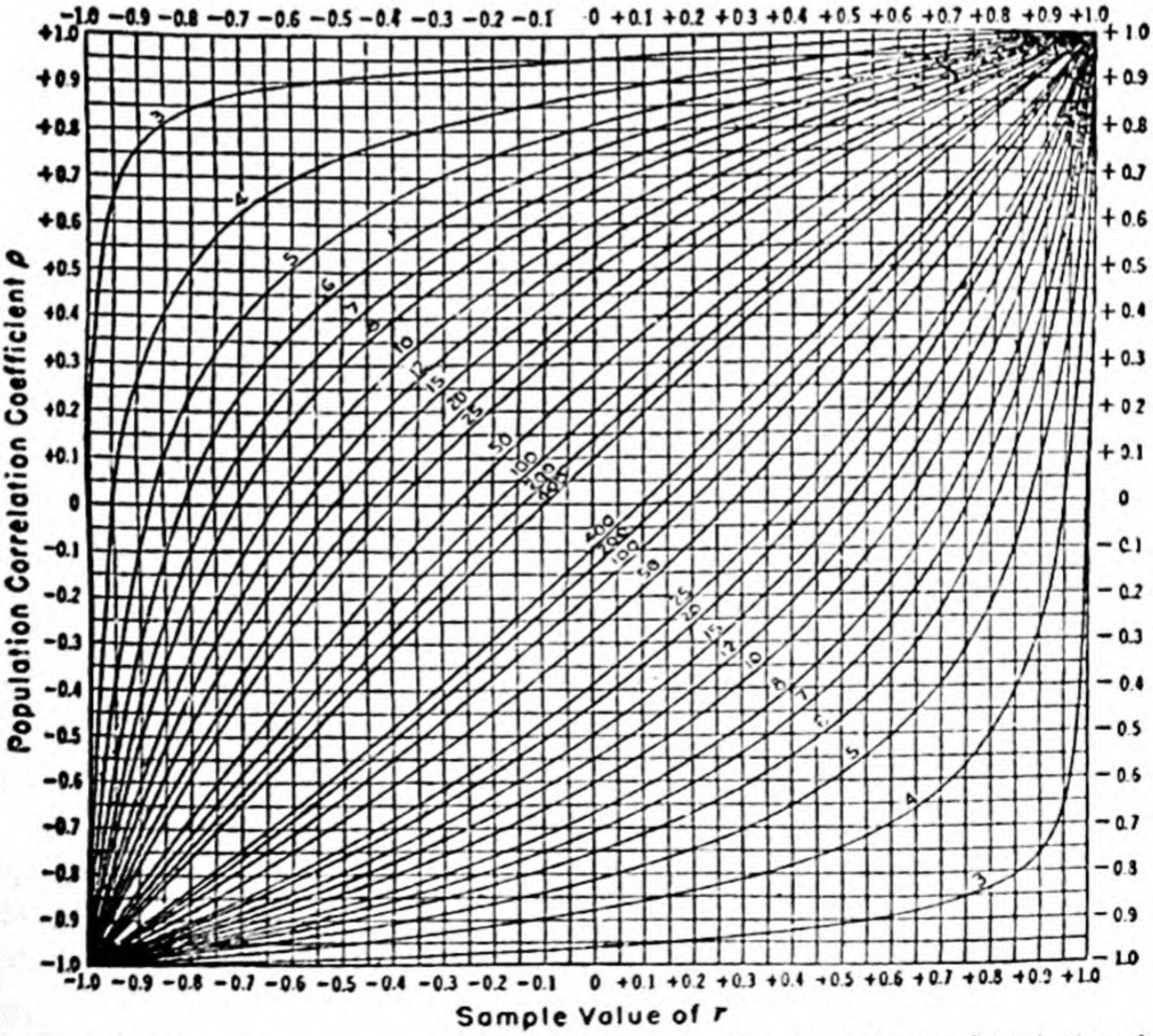
r	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.000	.010	.020	.030	.040	.050	.060	.070	.080	.090
.1	.100	.110	.121	.131	.141	.151	.161	.172	.181	.192
.2	.203	.214	.224	.234	.245	.256	.266	.277	.288	.299
.3	.309	.321	.332	.343	.354	.366	.377	.389	.400	.412
.4	.424	.436	.448	.460	.472	.485	.497	.510	.523	.536
.5	.549	.563	.577	.590	.604	.618	.633	648	.663	.678
.6	.693	.709	.725	.741	.758	.775	.793	.811	.829	.848
.7	.867	.887	.908	.929	.950	.973	.996	1.020	1.045	1.071
.8	1.099	1.127	1.157	1.188	1.221	1.256	1.293	1.333	1.376	1.422
.9	1.472	1.528	1.589	1.658	1.738	1.832	1.946	2.092	2.298	2.647

SOURCE: Elmer Mode, *Elements of Statistics*, 2nd ed. (New York: Prentice-Hall, Inc., 1951) Table N, p. 370.

Table VIII

Ninety-Five Per Cent Confidence Limits for the Coefficient of Correlation

(The numbers on the curves indicate sample size)



SOURCE: F. N. David, *Tables of the Ordinates and Probability Integral of the Distribution of the Correlation Coefficient in Small Samples* (London: The Biometrika Office, 1938), Chart II, by permission of Professor E. S. Pearson.

INDEX

a, 142, 143
Absolute value, 45
Acceptance region, 3, 91, 92, 93, 95, 97, 99
Alternative hypothesis (*See* Hypothesis, null and alternative)
Area:
  inside rectangles of histogram, 23, 24
  under frequency polygon, 24, 25
  under normal curve, 57, 58, 168
Arithmetic mean (*See* Mean)
Arsenian, Seth, 108
Association, 5, 12
  continuous variables, 139–157 (*See also* Correlation, coefficient of correlation)
  discrete variables:
    four-fold table analysis, 123–126, 129–134
    measure of degree, 134, 135
    testing by chi-square, 123–127
Average deviation (*See* Mean deviation)
Average (*See* Mean, Median, Mode)

b, slope, 142–143
  computed with deviations from a mean, 143, 144
  computed with observed values, 144, 152
Base for percentages, 12, 13, 14, 16, 17
Bettelheim and Janowitz, 2, 3, 6, 17
Bias, 80, 81, 82, 83 (*See also* Error)
Binomial distribution, 63–72
  application, 72
  in determining families at home, 63–68
  in toss of coins, 68–71
  mean, 110
  normal curve approximation, 71, 110
  p, probability of successful event, 64, 110
Binomial distribution (*Cont.*):
  q, probability of unsuccessful event, 64, 110
  standard deviation, 110
Burgess, E. W., and H. Locke, 18

Caplow, Theodore, 119
Categories:
  exhaustive, 10
  mutually exclusive, 10
  relevant, 10
Causation and correlation, 153–154
Centers, Richard, 137–139
Central value, measures of, 34–43 (*See also* Mean, Median, and Mode)
Chi-square distribution, 125–127
  correction factor, 127
  degrees of freedom, 125, 126, 127
  features, 125, 127
  table, 172
  test, 125, 134
Classification, 9
Class interval (*See* Interval, class)
Coefficient of correlation, 148
  interpretation, 149, 150
  product moment, grouped data, 152–154
  product moment, ungrouped data:
    computed with deviations from a mean, 149, 150
    computed with observed data, 152
  questions to ask about, 153, 155, 156, 157
  (*See also* Correlation)
  standard error, 150, 151
Coefficient of determination, 148, 149
Combinations, 65, 66, 67, 69, 70, 71, 86
Confidence intervals:
  effect of sample size on, 112, 113
  for coefficient of correlation, 176
  for mean, 89, 90, 99, 100
  for proportion, 112, 171
  for regression coefficient, 146
  table of, 112, 171

Confidence limits (*See* Confidence interval)
Context of discovery, 7
Context of justification, 7
Contingency table, 123 (*See also* Association, discrete variables)
Continuous variable, 14, 16, 56 (*See also* Association, continuous variables)
Control:
  through partial correlation, 157
  through subdivision, 130–134
Correction factor, chi square, 127
Correlation, 147–157
  and causation, 153–154
  and extrapolation, 155
  and linearity, 156
  and overlapping data, 155
  and regression, 150
  and skewness, 155
  negative, 149
  positive, 149
  spurious, 156, 157
  (*See also* Coefficient of correlation)
Crossley, 82
Cumulative frequency distribution, 25, 26, 54
Cumulative frequency polygon, 26

Dean, John, 120
Degrees of freedom:
  chi-square, 125, 126, 127
  t distribution, 106, 107
de Moivre, 55
Dependent variable, 140
Determination, coefficient of, 148, 149
Deviant cases, 6, 134
Discrete variable, 14, 63, 72 (*See also* Association, discrete variables)
Dispersion, measures of, 44–50 (*See also* Mean deviation, Range, Semi-interquartile range, Standard deviation, and Variance)
Distribution:
  binomial, 63–72 (*See also* Binomial distribution)
  chi-square, 125–134 (*See also* Chi-square distribution)
  frequency, 10–12, 54, 64
  normal, 51–61 (*See also* Normal curve)
  t, 105, 106, 169, 170
Dynamics of Prejudice, 2, 6

Edwards, Alba, 84
Equation, linear, 141, 142
Error:
  chance or sampling, 74, 79–81, 85
  in estimate, 141, 145
  of bias, 80, 81, 82, 83
  standard (*See* Standard error)
  type I, 90, 91, 92 (*See also* Level of significance)
  type II, 90, 92
Estimated values, 141
Exhaustive events, 65
Expected frequency, 124, 125, 127
Extrapolation, 155

Factorial, 67
Fearing, Franklin, 136
Festinger, L. Schachter, and K. Bach, 20
Finite universe:
  correction for, 101
  sampling from, 75
Fisher exact test, 127
Fourfold table, 123–126, 129–134 (*See also* Association, discrete variables)
Freedom, degrees of (*See* Degrees of freedom)
Frequency, 10
Frequency distribution, 10–12, 54, 64
Frequency expected, 124, 125, 127
Frequency observed, 123, 125
Frequency polygon, 24, 25, 39

Gallup, 82
Galton, 141
Graphic presentation, 21–32
  cumulative frequency polygon, 26
  frequency polygon, 24, 25, 39
  histogram, 23, 24, 27, 38, 51, 53, 64, 68, 71, 88
  pie chart, 21
  vertical bar graph, 21, 22
Guilford, 134

Haner and Meier, 82
Harlan, Howard, 108
Histogram, 23, 24, 27, 38, 51, 53, 64, 68, 71, 88
Hypothesis, exploratory, 2

Hypothesis, null and alternative, 3, 5, 90, 91, 92, 93
for chi-square, 124, 125, 129
for difference between two means, 102, 103, 105
for difference between two proportions, 117, 118, 119, 121, 122
for mean, 94, 95, 97, 98, 99
for proportion, 109, 110, 111
in *t* test, 106
Hypothesis, statistical, 3, 4

Independence, 64, 65, 68, 72, 123, 124, 130, 135
Independent variable, 17, 140, 146
Inductive process, 74
Interpolation, 38
Interval, class:
midpoint, 16, 35
number, 14, 15, 35
open-end, 15, 37
real limits for continuous variable, 16
size, 14, 15

Janowitz, Morris, 2, 3, 6, 17
Johnson, Charles, 18

Komorovsky, Mirra, 30

Lazarsfeld, Paul, 121
Lazarsfeld and Kendall, 136
Least squares, line of, 141, 143, 144
Level of significance, 91, 94, 95, 97, 103, 109, 111 (*See also* Error, type I)
Levinson, Daniel, 50
Limits (*See* Confidence interval and Interval, class)
Linear correlation (*See* Coefficient of correlation, Correlation)
Linear equation, 141, 142
Linear regression, 139–147
X on *Y*, 146
Y on *X*, 141, 144, 145, 146
Linear relationship, 140, 147
Line of best fit, 141, 143, 144 (*See also* Linear regression)

Maximum likelihood test, 127
Mean:
arithmetic, from grouped data, 35, 36, 37, 53
Mean (*Cont.*):
arithmetic, from ungrouped data, 34, 35, 44
arithmetic, properties of, 40
arithmetic, use of, 40, 41
confidence interval for, 89, 90, 99, 100
of binomial distribution, 110
of normal distribution, 55
sampling distribution of, 85–89 (*See also* Sampling distribution of means)
standard error of, 87, 97, 98, 105
statistical test for, 94–99
Mean deviation, 44, 45
Median, 27, 37, 48, 49
assumptions with grouped data, 39
from grouped data, 38
from ungrouped data, 38
use of, 40, 41
Merton, Robert, 29
Mode, 39
relation to mean and median, 39
use of, 40, 41
Mosteller, F., 154, 158
Mutually exclusive events, 64

Normal curve, 51–61
and binomial distribution, 71, 110
and sampling distribution, 88
area under, 57, 58
description, 55, 56
fitting of, to histogram, 56, 57
for distribution of test scores, 60, 61
for sociological data, 60, 61
table of, 59, 168
Normal curve graph paper, 53, 55
Null hypothesis (*See* Hypothesis, null and alternative)

Observational data, 1
Observed frequency, 123, 125
Observed values, 141
Open-end interval, 15, 37
Operational concepts, 4
Overlapping data, 155

p, probability that event will occur, 64, 110
Parameter, 85
Pearson, 141

Pearson-product-moment coefficient of correlation (*See* Coefficient of correlation, Correlation)
Percentiles, 27
Peters and Van Voorhis, 134
Pie chart, 21
Polygon:
  cumulative frequency, 26
  frequency, 24, 25, 39
Population, definition of, 74
Positive correlation, 149
Power of a test, 93, 94
Probability, definition, 63
Product-moment coefficient of correlation (*See* Coefficient of correlation, Correlation)
Proportion:
  confidence interval, 112, 171
  sampling distribution, 109–110
  standard error, 110, 111
  statistical test, 111, 112

q, probability that event will not occur, 64, 110
Quartiles, 27, 48, 49
Quota sampling, 82–83

r (*See* Coefficient of correlation)
Random numbers, 75, 77, 109, 174
Random sample (*See* Sampling, probability)
Random selection, 75
Range, 44, 45
Raw score, 57, 58
Region of acceptance, 3, 91, 92, 93, 95, 97, 99
Region of rejection, 3, 5, 91, 92, 93, 94, 95, 97, 99, 103, 107, 117
Regression:
  and correlation, 150
  coefficient of, 142–143 (*See also* b, slope)
Reichenbach, 6
Rejection region, 3, 5, 91, 92, 93, 94, 95, 97, 99, 103, 107, 117
Relative frequency, 11, 24
Reliability, 5, 157
Representative, 6
Roper, 82

s, standard deviation of sample (*See* Standard deviation)
Sample, 74, 85
Sample, random (*See* Sampling, probability)
Sample size:
  determination of, 114, 115
  effect on confidence interval, 113
Sampling:
  probability, 74–79, 82
    cluster, 77, 78
    combination of stratified and cluster, 78, 79
    comparison with quota sampling, 82
    difficulty with, 81
    simple random, 75–79
    stratified random, 76–78
    systematic, 77
  quota, 82–83
Sampling distribution, definition, 85–86 (*See also* Chi-square distribution; Sampling distribution of means; Proportion, sampling distribution; t distribution)
Sampling distribution of means, 85–89
  approximation to a normal curve, 88, 90, 91, 97
  mean of, 86, 87
  standard deviation of, 87
  variance of, 87, 88
Sampling error (*See* Error: chance or sampling)
Scatter diagram, 140, 156, 157
Scientific method, 1, 2
Score, standard (*See* Standard score)
Semi-interquartile range, 48, 49
Sex ratio, 12
Sigma, σ, standard deviation of universe (*See* Standard deviation)
Significance, level of, 91, 94, 95, 97, 103, 109, 111 (*See also* Error, type I)
Skewness, skewed distribution, 27, 37, 39, 155
Slope of line, 142 (*See also* b, slope)
Spurious relationship, 131
Squares, table of, 166
Standard deviation:
  around regression line, 145, 146
  formulas, grouped data:
    using deviations from guessed mean, 47, 52
    using deviations from mean, 96
    using observed values, 47

Standard deviation (*Cont.*):
formulas, ungrouped data:
using deviations from mean, 44, 46
using observed values, 46
measure of distance under normal curve, 48, 55
Standard error:
correction for finite sampling, 101
of correlation coefficient, 150, 151
of difference between means, 104
of difference between proportions, 118, 122
of estimate, 145, 146
of mean, 87, 97, 98, 105
of proportion, 110, 111
of Z transformation, 151, 152
size as compared with standard deviation, 100, 101
size of, 115, 116
variation with change in p, 113
variation with size of sample, 89
Standard score:
converting from raw score, 57, 58
for correlation coefficient, 151
for difference between means, 105
for difference between proportions, 118, 122
for means, 95, 97, 98, 99
for proportions, 111, 116
for Z values, 151, 152
Statistic, 85
Statistical inference:
estimating universe parameter (*See* Confidence interval)
testing statistical hypotheses (*See* Acceptance region: Hypothesis, null and alternative; Level of significance; Rejection region; Standard error; Error, type I and type II; Standard score)
Statistical test (*See* Statistical inference for references)
Statistics, 1
Stouffer and Lazarsfeld, 121
Stouffer, Samuel, 31, 32, 121
Strata, 76, 78
Stratification, 76, 77, 78
"Student," 106
Student-t distribution, 105, 106
Subdivision, 130–134 (*See also* Strata)
Symmetry, symmetrical distribution, 27, 37, 39, 48, 49
T^2, 134, 135
t distribution, 105, 106
t table, 169, 170
t test, 105, 106, 107
Table:
construction, 12
one-way, 10
two-way, 11
Tests, statistical (*See* Statistical inference for references)
Theoretical distribution (*See* Binomial, Chi-square, Normal curve, and t distribution)
Transformation, Z, 151, 176
Type I error, 90, 91
Type II error, 90, 92

Universe, 2, 74, 85
finite, 75, 101 (*See also* Finite universe)

Validity, 5, 157
Variable, 14
antecedent, 131, 133
continuous, 14
dependent, 140, 141, 150
discrete, 14, 63, 72
nonqualitative, 14
qualitative, 14
independent, 17, 140, 141, 146, 150
intervening, 133
Variance, 44, 45, 46, 86
around mean, 148
around regression line, 148
explained, 148
unexplained, 148
Vertical bar graph, 21, 22

Weighted mean, 40
Whyte, William F., 32
Wilner, D., and F. Fearing, 136
Wirth, Louis, and H. Goldhamer, 19

z score, 95, 97, 98, 99, 105, 106, 111, 116, 118, 122 (*See also* Standard score)
Z transformation, 151, 176
Zero order correlation (*See* Coefficient of correlation, Correlation)